# ESSAYS IN BEHAVIORAL AND HEALTH ECONOMICS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Lawrence Jung Kee Jin

May 2018

# ESSAYS IN BEHAVIORAL AND HEALTH ECONOMICS

Lawrence Jung Kee Jin, Ph. D.

Cornell University 2018

This dissertation consists of three chapters. The first two chapters provide empirical evidence of hot-hand bias in two novel field settings: dart players' strategic choices, and physicians' decisions during childbirth. The "hot hand" refers to the notion that a person can enter a state in which her probability of success becomes higher than normal. Regardless of whether the person actually has a hot hand, the "hot-hand bias" is when the person has an exaggerated belief about the hot hand. In Chapter 1, I collect data of professional dart players from the 2016 World Darts Championship. I find that players are significantly more likely to hit after a successful shot, implying that players have a hot hand. Based on a precise estimate of the hot hand, I calculate the optimal strategy of a profit-maximizing dart player. I find that dart players are much more willing to take risks after a successful shot than what I calculate to be optimal, consistent with hot-hand bias.

In Chapter 2, I utilize 1.3 million hospital admissions for childbirth in New York State over 2010-2015. I find no evidence that physicians have a hot hand when performing obstetrical procedures. In the absence of hot hand, physicians are still

2% more likely to perform a C-section after a previous successful C-section. My empirical model includes physician fixed effects and a large set of patient conditions that proxy for when a C-section is likely to maximize patient welfare. Robustness checks provide additional evidence consistent with decision-makers having hot-hand bias. Assuming that the identified 2% increase in the C-section rate is unwarranted, the estimated health-care cost is $65 million per year in the US.

Chapter 3 is joint work with Nicolas Ziebarth. We investigate the relationship between sleep and health using a census of 160 million hospital admissions from Germany and 3.4 million survey responses from the US over one decade. We exploit the exogenous extension of sleep when daylight saving time ends: setting clocks back by one hour in the fall significantly extends night's sleep and reduces self-reported tiredness for four days following the time shift. In turn, we find that self-reported health improves and hospital admissions decrease significantly for about four days.

## BIOGRAPHICAL SKETCH

Prior to Ph.D., Lawrence Jin received a bachelor's degree in mathematics and economics from Cornell University, and spent one year working at a surgery department in Hong Kong. His research interests lie at the intersection of behavioral and health economics. His current research focuses on areas in which behavioral economics can help us better understand how people make health-related decisions, often with an eye towards the implications for public health policy.

To my family, and Dom

# ACKNOWLEDGEMENTS

I have received support and encouragement from a great number of individuals. I would first like to express my sincere gratitude to my adviser, Donald Kenkel. He has been my mentor since the undergraduate years, and his thoughtful and warm guidance has made Ph.D. a truly rewarding journey. I am also forever indebted to my dissertation committee members, Daniel Benjamin and Tatiana Homonoff, for their support and guidance. They continue to inspire me and help me grow as a scholar. I am also extremely grateful to Bryant Kim, Ted O'Donoghue, and Nicolas Ziebarth for their kind mentorship throughout the years.

I feel fortunate to have received so much love and support from family and friends. I have incredible parents who are supportive, encouraging, and sometimes far too excited about the things I do. I am grateful to Jen, Olivia and Grandma for always making me feel at home even when we are seven thousand miles away. Finally, I thank all my friends – especially Dayoung for being my soul mate and providing stability and balance to my Ph.D. life.

# Contents

**1 Don't Overshoot: Evidence of Hot-Hand Bias from World Darts Championship** **1**

1.1 Introduction . . . . . 1

1.2 Institutional Details and the Data . . . . . 4

1.3 Empirical Framework and Theory . . . . . 8

1.3.1 Testing the Hot Hand . . . . . 8

1.3.2 Testing the Hot-Hand Bias . . . . . 9

1.4 Results . . . . . 12

1.4.1 Test for the Hot Hand . . . . . 12

1.4.2 Test for the Hot-Hand Bias . . . . . 13

1.4.3 Robustness Checks and Heterogeneity . . . . . 15

1.5 Conclusion . . . . . 20

**2 Evidence of Hot-Hand Bias in Medical Decision-Making** **21**

2.1 Introduction . . . . . 21

2.2 Institutional Details and the Data . . . . . 24

2.3 Empirical Specification . . . . . 28
2.4 Results . . . . . 30
2.4.1 Test for the Hot-Hand . . . . . 30
2.4.2 Test for the Hot-Hand Bias . . . . . 30
2.4.3 Robustness Checks . . . . . 35
2.4.4 Discussion of Mechanisms . . . . . 36
2.5 Conclusion . . . . . 38

**3 Sleep and Health: Evidence from Daylight Saving Time (with Nicolas Ziebarth) 40**

3.1 Introduction . . . . . 40
3.2 Datasets . . . . . 43
3.2.1 The US Behavioral Risk Factor Surveillance System (BRFSS) 43
3.2.2 German Hospital Admissions Census . . . . . 46
3.3 Empirical Specification . . . . . 48
3.3.1 Main Specification . . . . . 48
3.3.2 Identification . . . . . 49
3.4 Results . . . . . 51
3.4.1 The Effect of Time Shift on Sleep . . . . . 51
3.4.2 The Effect of Time Shift on Hospital Admissions . . . . . 54
3.4.3 The Effect of Time Shift on Self-Reported Health . . . . . 57
3.4.4 Could Alternative Mechanisms Explain the Health Effects? . 60
3.5 Conclusion . . . . . 61

# Chapter 1

# Don't Overshoot: Evidence of Hot-Hand Bias from World Darts Championship

## 1.1 Introduction

There is a widespread belief that athletes have "hot hands," i.e. they can sometimes enter a "hot" state in which the probability of success becomes higher than normal. Earlier evidence has suggested that the hot hand is a fallacy, based on empirical evidence that performance outcomes appear not to be serially correlated (Gilovich, Vallone, and Tversky, 1985).[1] However, recent studies have found significant evidence of hot hand in sports upon correcting for a statistical bias in previous

[1]Raab, Gula, and Gigerenzer (2012) provide a review of the hot-hand literature in sports.

measures of conditional probabilities (Miller and Sanjurjo, 2015) and after controlling for endogenous responses of opponents (Green and Zwiebel, 2017; Miller and Sanjurjo, 2014).

In the *presence* of hot hand, it is challenging to test whether people have correct perception of the hot hand. Camerer (1989) studies the sports betting market and finds that people overbet on teams that are on a winning streak. This is consistent with "hot-hand bias," i.e. an exaggerated belief in the hot hand. The stakes would potentially be higher if athletes themselves are subject to the hot-hand bias, which can distort their behavior during the game. Green and Zwiebel (2017) find that professional baseball pitchers overreact to batters who have been hot in their previous five at-bats, and walk them more than can be justified by the batters' true performance.

This paper provides first field evidence of professional athletes overreacting to their *own* hot hand. Specifically, I study professional dart players' performances during a large international tournament. A dart player chooses from several possible targets on the dart board, each with varying risks and rewards. Where their shots land reveal whether they chose a risky strategy or a safe one. Because the goal of the game is to hit as many points as possible (in the early game), I can calculate the point-maximizing optimal strategy based on the players' true hot hand. By comparing the players' strategies to the calculated optimal strategy, I can test whether they overshoot when hot, which would be consistent with hot-hand bias.

The data are collected from the 2016 World Darts Championship, where the world's best dart players competed for a total prize pool of £1.5 million (US $2

million). First, I find that professional dart players have a large hot hand. After a previous successful hit, players are 18% more likely to hit the intended target again ($p < 0.001$). Next, I test whether the players have correct perception of their hot hand. I calculate the point-maximizing optimal target choices based on the players' true hot-hand magnitudes. Compared to the calculated optimal strategy, I find that players are 8% more risk-taking after a previous success ($p < 0.001$). The evidence suggests that professional dart players suffer from significant hot-hand bias. Robustness checks rule out alternative explanations, such as a possible tendency for players to become more risk-taking when they are ahead in the game. The cost of this strategic error is large, at about $960 per match in lost earnings.

This paper is one of the first studies to show that the hot-hand bias distorts behavior in an economically significant way. This has traditionally been one of the challenges of the hot-hand literature. For example, the forecasting error due to hot-hand bias in the basketball betting market is only slight (Camerer, 1989). In the study by Green and Zwiebel (2017), it is hard to check whether or not a pitcher's overrection to the batters' hot hand incurs a tangible cost to the team's performance. In baseball, there are too many factors that can counteract a pitcher's mistake, such as referees being more generous to the pitcher after a walk. Finally, studies have shown consistent evidence of hot-hand bias among gamblers, such as when playing the roulette in casinos (Croson and Sundali, 2005), buying lotteries from "lucky stores" (Guryan and Kearney, 2008), and picking lottery numbers that have won frequently in the recent past (Suetens, Galbo-Jørgensen, and Tyran, 2015). However, these "mistakes" are not costly because the outcomes are random and

therefore independent of the gambler's decisions.[2]

The paper is organized as follows. Section 1.2 provides details about the game and the data, and Section 1.3 describes the empirical framework. The results are presented in Section 1.4 and Section 1.5 concludes.

## 1.2 Institutional Details and the Data

Darts is a two-player game, and each player starts a *leg* with 501 points. Players take turns to throw three darts in successsion, and the three darts are removed from the board only after the turn is over. The dart board is split into 20 areas, each with a number between 1 and 20 (Figure 1.2.1). This is the baseline number of points that a player will receive by hitting the area with a dart. In addition, there are two rings on the board that will either double or triple the baseline points. The red circle in the middle is the Bull's Eye, worth 50 points, and the green area surrounding it is worth 25 points. The maximum points attainable with a shot is 60 points, by hitting a Triple-20 ("60" henceforth). The next highest target is a Triple-19 ("57"), followed by a Triple-18 ("54").

The goal of the game is to accumulate exactly 501 points before the opponent does, bringing one's score down from 501 to zero. The last throw has to be a Double or a Bull's Eye. If a player hits more points than required to reach zero, the player "busts" and the score returns to the score at the start of that turn. A *set* is won when a player wins three legs, and the match is won by the first player to win three

[2]Suetens, Galbo-Jørgensen, and Tyran (2015) find that picking "hot" lottery numbers actually incurs a small cost because more people tend to pick these numbers, so when they win, they have to split the winnings across more people.

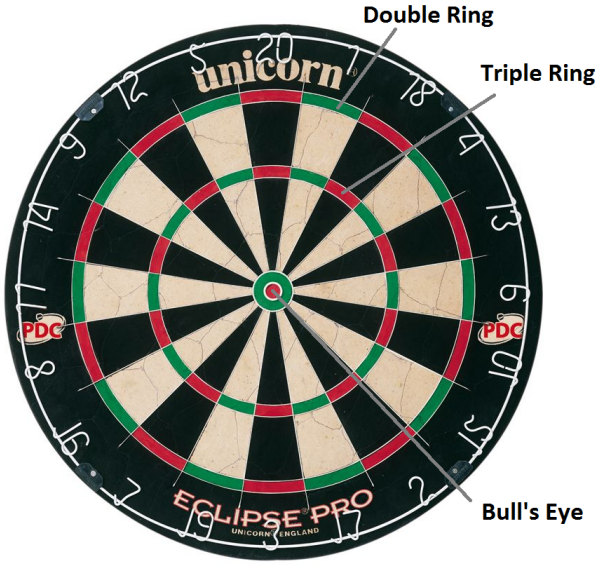



Figure 1.2.1: Dart Board

(round 1), four (rounds 2 & 3), five (quarterfinals), six (semifinals), or seven (final) sets.

The optimal strategy is to first try and hit as many points as possible to bring the score down quickly. As the score approaches zero, players must then consider the optimal paths to "close out" the game. To eliminate the complexities of the endgame, I restrict the sample to the first nine throws of a leg, where players unambiguously try to hit as many points as possible.

The data are from the 2016 World Darts Championship, which is a prestigious international darts tournament held annually in London. The tournament began in December 17, 2015 and concluded in January 3, 2016. The total prize pool was £1.5 million (US $1.97 million), with £300,000 ($394,000) given to the eventual champion

Gary Anderson.[3]

Four research assistants were recruited to watch 31 tournament videos on Youtube, starting from Round 2 to the Grand Final. Using a computer program, the research assistants recorded the coordinates of 21,188 shots played by 32 players in the tournament. After restricting the sample to the first nine shots of a leg, the final sample size is 11,810 shots.

Figure 1.2.2 shows a heat map of the sample. 82% of shots landed near 60, and these are coded as being aimed at 60. 15% of shots landed near 57, 3% landed near 54, and the remaining 1% of darts are scattered around 51 and the Bull's Eye. The intended target of a shot is determined by the closest distance to the following five targets: 60, 57, 54, 51, and 50.

Table 1.1 shows summary statistics. On average, players hit the intended target 41% of the time. The hit rate is slightly lower at the start of a turn, at 36%, but the hit rate improves for the 2nd and 3rd shots of a turn. This may be because the first shot is made after a short break while the opponent completes his turn.

Players are significantly more likely to hit after a hit, compared to after a miss (46% vs 37%, $p < 0.01$), consistent with players having a hot hand. This difference is not significant when restricted to the first shots of a turn, suggesting that a time delay between turns might be eliminating the hot hand. However, there are strong indications that hot hand exists when we look at 2nd and 3rd shots of a turn.

On average, 82% of the shots in my sample are aimed at target 60. The first shots of a turn are almost always aimed at 60, but the frequency drops to 79% among 2nd

[3]The exchange rates as of November 10, 2017

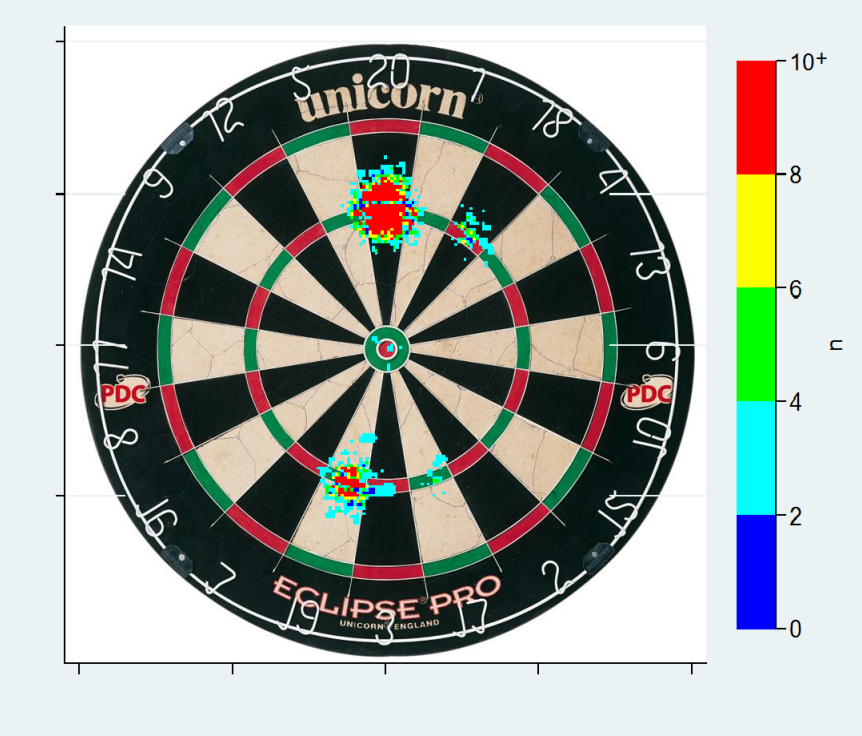



Figure 1.2.2: Heat Map

Table 1.1: Summary Statistics (Darts)

| | Mean | $HIT_{t-1} = 0$ | $HIT_{t-1} = 1$ | $p$-value |
|---|---|---|---|---|
| Hit | 0.41 | 0.37 | 0.46 | $< 0.01$ |
| 1st Shot | 0.36 | 0.35 | 0.38 | 0.11 |
| 2nd Shot | 0.42 | 0.38 | 0.50 | $< 0.01$ |
| 3rd Shot | 0.44 | 0.38 | 0.51 | $< 0.01$ |
| | | | | |
| Shoot at 60 | 0.82 | 0.77 | 0.89 | $< 0.01$ |
| 1st Shot | 0.97 | 0.97 | 0.97 | 0.39 |
| 2nd Shot | 0.79 | 0.70 | 0.93 | $< 0.01$ |
| 3rd Shot | 0.67 | 0.61 | 0.75 | $< 0.01$ |
| # Players | | 32 | | |
| # Matches | | 31 | | |
| # Shots | | 11,810 | | |

Notes: This table presents summary statistics of the darts data collected from the 2016 World Darts Championship. The data include 31 matches from Round 2 to the Grand Final.

shots, and to 70% for 3rd shots. This is because after the player shoots at 60, the dart remains on the board and partially blocks the target, making it slightly harder to hit 60 again. We also see that players are much more risk-taking after a previous hit. For the 2nd shot, players will almost always attempt a 60 again after a hit, significantly more than if the prior shot was a miss (93% vs 72%, $p < 0.01$). Similarly, for the 3rd shot, players shoot at 60 77% of the time after a hit, compared to 64% after a miss ($p < 0.01$). The results are qualitatively consistent with an expected payoff maximizer: when a player has made a previous shot, he is more likely to hit the target again, and therefore should attempt the risky shot more often.

## 1.3 Empirical Framework and Theory

### 1.3.1 Testing the Hot Hand

The general empirical specification for testing the hot hand is shown below:

$$S_{i,t} = \alpha_0 + \alpha_1 S_{i,t-1} + \alpha_2 X_{i,t} + \epsilon_{i,t}$$

$S_{i,t}$ is an indicator for a successful outcome by decision-maker $i$'s performance at time $t$. In darts, the success is defined as whether a player's shot has hit the intended target. $X_{i,t}$ is a vector of control variables that need to ensure that the explanatory variable, $S_{i,t-1}$, is conditionally random.

$\alpha_1$ is interpreted as the change in probability of success when the decision-maker has previously succeeded. $\alpha_1 > 0$ is evidence in favor of the existence of a hot hand.

$\alpha_1 = 0$ would imply that there is no hot hand. $\alpha_1 < 0$ implies outcome reversals, where the chance of success decreases following a previous success.[4]

### 1.3.2 Testing the Hot-Hand Bias

Dart players have an objective to try and hit as many points as possible (in my analysis sample; see Section 1.2). This means a player's decision to choose a particular target, say 60, should only depend on its expected payoff (i.e. number of points) relative to the next-best option. For example, in a world without measurement error, the player would choose 60 if and only if the expected payoff of choosing 60 is higher than the next-best option. With measurement error, the player should still respond positively to the expected payoff of 60 relative to other options.

The key idea is that if players are rational agents who want to maximize the chance of winning the match, their decision to choose 60 should *only* depend on the expected payoff of 60 relative to the next-best option. Put differently, after controlling for the relative expected payoff of 60, the players' decision to choose 60 should no longer be influenced by other factors such as a prior success. If the decision is positively (negatively) affected by the prior success, it suggests an overreaction (underreaction) to the hot hand.

I begin with a simple model of a dart player. Let $i$ index the player, $j$ index the possible targets on the board, and $s$ index the state of the world. Player $i$'s utility of choosing target $j$ in state $s$ is given by the expected payoff of chooding target $j$ conditional on state $s$, denoted by $V_{i,j,s}$, and an unobserved error term:

[4]One possible explanation is that the decision-maker puts less effort after a previous success.

$$U_{i,j,s} = V_{i,j,s} + \epsilon_{i,j,s}$$

The player chooses a target that maximizes the utility, as shown:

$$U_{i,j,s} \geq U_{i,k,s}, \forall k \neq j$$

The expected payoff of choosing a target is approximated by the observed average number of points scored by the players when they chose that target conditional on the observable state of the world. In the baseline specification, I consider 2 possible targets (60 or 57), and 48 possible states of the world for each target, as shown:

$$\hat{U}_{i,j,s} = \sigma_{j,h_1,a_1,h_2,a_2,n}$$

$\hat{U}_{i,j,s}$ denotes the expected payoff (or utility) of player $i$ choosing target $j$ in state $s$. $j$ denotes two possible targets, 60 or 57. The reason for considering only two possible targets is that there are not enough observations of players choosing targets other than 60 and 57. $h_1$ denotes a "hot" state of the player, defined to be 1 if the previous shot was a success, and 0 otherwise. $a_1$ is an indicator for whether the previous shot was aimed at 60 or not. I also consider $h_2$ and $a_2$ that correspond to whether two shots ago was a success, and whether that shot was aimed at 60, respectively. Finally, $n$ denotes the three possible stages of a turn: 1st shot, 2nd shot, or 3rd shot.

This generates 96 possible expected payoffs: 2 possible targets in 48 possible states of the world (2 ($j$) x 2 ($h_1$) x 2 ($a_1$) x 2 ($h_2$) x 2 ($a_2$) x 3 ($n$)).

What matters when choosing a target is its expected payoff relative to the second-best option. Let $\hat{X}_{i,j,s}$ denote the relative expected payoff of choosing target $j$ in state $s$, obtained by subtracting the estimated expected payoff of the next best option $k \neq j$, as shown:

$$\hat{X}_{i,j,s} = \hat{U}_{i,j,h} - \hat{U}_{i,k,h}$$

A rational agent's decision to choose the risky target of 60 would be determined by its relative expected payoff. I therefore estimate the following OLS:

$$Y_{i,t} = \sigma_i + \beta_1 \hat{X}_{i,t} + \beta_2 Hit_{i,t-1} + \epsilon_{i,t}$$

$Y_{i,t}$ is the decision variable that equals 1 if player $i$ chooses target 60 at time $t$. $\hat{X}_{i,t}$ is the relative expected payoff of target 60 at time $t$ estimated in the first-stage. Because it is estimated by conditioning on a previous success, it incorporates the benefits of having a hot hand that makes 60 a more attractive target. $Hit_{i,t-1}$ is an empirical proxy for being in a hot state, defined to be 1 if the previous shot at $t-1$ was a successful hit. The standard errors are clustered at the player level.

$\beta_2$ is the coefficient of interest, and it is interpreted as the change in probability of choosing target 60 when the player has previously hit, conditional on the relative expected payoff of 60. $\beta_2 = 0$ would be consistent with a rational expected-payoff maximizer. On the other hand, $\beta_2 < 0$ would be consistent with players underreacting to having a hot hand, and $\beta_2 > 0$ might suggest an overreaction.

Table 1.2: Test for the Hot-Hand (Darts)

| | (1) | (2) |
|---|---|---|
| | DV = Hit | DV = Points Scored |
| $Hit_{t-1}$ | 0.075*** | 3.365*** |
| | (0.010) | (0.427) |
| Player fixed effects | Y | Y |
| Constant | 0.384*** | 33.750*** |
| | (0.004) | (0.173) |
| Obs | 11,810 | 11,810 |

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. $Hit_{t-1}$ is an indicator that equals 1 if the dart player's previous shot has hit the intended target. In Column 1, the dependent variable is whether the current shot has hit the intended target. The regression tests whether a previous hit is predictive of current hit, which would indicate that professional dart players have a hot hand. In Column 2, the dependent variable is the number of points that a player has scored with the current shot. The regression tests whether a previous hit improves the average number of points that a player scores with the current shot. Standard errors are clustered at the player level.

## 1.4 Results

### 1.4.1 Test for the Hot Hand

There is strong evidence that hot hand exists in professional darts performances. In Table 1.2, Column 1, I find that the probability of hitting the intended target improves by 7.5 percentage points (ppt) following a successful hit ($p < 0.001$). This is an 18% improvement over the average hit rate of 41%. Column 2 shows that, on average, players score 3.4 more points when in a hot state. The results show that, indeed, the hot hand is not a fallacy, consistent with recent evidence from professional basketball and baseball (Miller and Sanjurjo, 2014, 2015; Green and Zwiebel, 2017).

Table 1.3: Test for the Hot-Hand Bias (Darts)

| Dep Var = Shoot at 60 | (1) | (2) | (3) |
|---|---|---|---|
| $Hit_{t-1}$ | 0.140*** | 0.065*** | 0.066*** |
| | (0.016) | (0.012) | (0.012) |
| $\hat{X}_{i,t}$: The relative expected payoff of 60 | | 0.019*** | 0.002* |
| | | (0.002) | (0.001) |
| Indicator: 60 has the highest expected payoff | | | 0.447*** |
| | | | (0.037) |
| Player fixed effects | Y | Y | Y |
| Constant | 0.731*** | 0.712*** | 0.381*** |
| | (0.006) | (0.009) | (0.034) |
| Obs | 11,810 | 11,810 | 11,810 |

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The table tests whether the decision to choose the maximum-possible target is related to having a hot hand. The dependent variable is an indicator that equals 1 if the player shoots at the maximum-possible target of 60. $Hit_{t-1}$ is an indicator for whether the dart player's previous shot has hit the intended target, which is a proxy for having a hot hand. $\hat{X}_{i,t}$ is the estimated expected payoff of target 60, minus the estimated expected payoff of the next-highest target of 57, based on the current state of the world (including whether the player is in a hot state or not). Column 3 also includes an indicator that equals 1 if $\hat{X}_{i,t} > 0$, i.e. if target 60 is calculated to be the optimal choice. Standard errors are clustered at the player level.

### 1.4.2 Test for the Hot-Hand Bias

Table 1.3 shows the estimated effects of a previous hit on the players' decision to choose target 60. In Column 1, without controlling for the relative expected payoffs, players are 14 ppts more likely to choose target 60 after a successful hit. This is qualitatively consistent with what a rational agent would do. Since having a hot hand significantly increases the probability of making the next shot, it would be payoff-maximizing to choose the high-reward target more often.

In Column 2, I control for the relative expected payoff of target 60, which incorporates the payoff benefits of having a hot hand. A five-point increase in the

relative expected payoff of 60 increases the likelihood of choosing 60 by 9.5 ppts ($p < 0.001$). The fact that the point estimate is positive and significant suggests that players are, in general, responding to the expected payoffs. However, they still respond significantly to a previous success. The new point estimate is smaller, at 6.5 ppts ($p < 0.001$). The fact that it is smaller implies that much of the large response found in Column 1 is due to the payoff benefits of having a hot hand. The new estimate is still significantly positive, suggesting that players are overshooting when they are hot.

Column 3 includes an indicator that equals 1 if target 60 has the highest expected payoff, that is, 60 is calculated to be the optimal choice. Players are 45 ppts more likely to choose 60 when 60 is calculated to be optimal ($p < 0.001$). There is now only a marginally significant positive response to the relative expected payoff of 60. Again, evidence suggests that players are generally payoff-maximizing. The coefficient for hot-hand bias remains significant at 6.6 ppts ($p < 0.001$). Taken together, the results suggest that players are much too risk-taking after a previous success than what I calculate to be optimum.

The regression estimates from Table 1.3 show the difference in risk-taking behavior between having made the previous shot compared to having missed the previous shot. The estimates suggest that this difference is larger than what I calculate to be optimum, but these estimates do not reveal *when* players are making mistakes. That is, are players making mistakes when they are "hot", or when they are "cold", or both?

Table 1.4 compares the observed versus optimal frequencies of choosing target

Table 1.4: Comparison Between Observed and Optimal Strategy (Darts)

| | Cold ($Hit_{t-1} = 0$) | | Hot ($Hit_{t-1} = 1$) | |
|---|---|---|---|---|
| | Observed Frequency of 60 (%) | Optimal Frequency of 60 | Observed Frequency of 60 | Optimal Frequency of 60 |
| 1st Shot | 96.8 | 100 | 97.1 | 100 |
| 2nd Shot | 70.4 | 74.2 | 93.1 | 97.9 |
| 3rd Shot | 61.1 | 78.6 | 74.6 | 79.5 |

Notes: This table compares the observed frequency of players choosing the maximum-possible target 60 vs. the calculated optimal frequency of choosing 60, when the player is hot or cold, and across the three stages of a turn.

60, across two dimensions: whether or not a player has a hot hand, and the nth shot of a turn. Interestingly, when players have a hot hand, they seem to be making choices that are generally consistent with what I calculate to be optimal. However, when a player has previous missed a shot, they seem to be appropriately risk-taking for the first two shots of a turn, but are not choosing 60 frequently enough for the 3rd shots. This perhaps suggests a *cold-hand* bias: after missing the previous shot, players underestimate their ability to hit the risky target, and as a result turn to lower-reward targets too often.[5]

### 1.4.3 Robustness Checks and Heterogeneity

In this section I consider a few alternative explanations of the results. I then test for heterogeneity of the hot hand and hot-hand bias across players.

When a player has successfully hit the previous shot, the player is in a better state of the game than if the previous shot was unsuccessful. It is therefore possible

[5]One may wonder if players are too conservative at the baseline, and have exaggerated hot hand beliefs after making a successful shot. This appears to be inconsistent with the evidence. For example, looking at the first shots of a turn in Table 4, players appear to be choosing the risky target 97% of the time even after having missed the previous shot.

that a response to a previous success is driven not by the hot-hand bias, but instead by a tendency to become more risk-taking when the player is ahead in the game.

In Table 1.5, Column 1 replicates the main results (Table 1.4, Column 3). In Column 2, I include the score difference, that is, the player's current score minus the opponent's score. A negative coefficient suggests that players are in fact *less* risk-taking when they are ahead. Being ahead by 100 points is associated with a 3 ppt decrease in the frequency of shooting at target 60 ($p < 0.001$). After controlling for the score difference, the estimated hot hand response is 7.6 ppts ($p < 0.001$).

Another potential concern is that a response to a previous success may be driven by the players' desire to hit a maximum of 180 points in a turn. The crowd cheers and the commentators go wild when a player hits three 60's in a row in a turn. In Column 3, I include an indicator that equals 1 if the player is facing the 3rd shot of a turn and his two previous shots have each hit 60. Players are 2.2 ppts more likely to aim at 60 again ($p = 0.098$). The point estimate for the hot-hand bias is 6.3 ppts ($p < 0.001$). Finally, Column 4 includes both control variables, and players are still more risk-taking after a successful shot than what is calculated to be optimal.

Figure 1.4.1 shows player-specific estimates of the hot hand for the 32 players in the sample. The players are ordered by their world rankings at the time of the tournament, starting with the world's best player, van Gerwen, on the left, and the lowest ranked van de Bergh (60th in the world) on the right. The blue bars show the estimated hot hand for each player, with 95% confidence intervals and a linear trend (dotted line). Note that the confidence intervals are generally larger for lower-ranked players on the right because they tend to drop out earlier in the tournament

Table 1.5: Robustness Checks (Darts)

| Dep Var = Shoot at 60 | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| $Hit_{t-1}$ | 0.066*** | 0.076*** | 0.063*** | 0.072*** |
| | (0.012) | (0.010) | (0.012) | (0.012) |
| Score difference with opponent (x 100) | | -0.030*** | | -0.031*** |
| | | (0.004) | | (0.014) |
| Chance for 180 | | | 0.022* | 0.040*** |
| | | | (0.013) | (0.014) |
| $\hat{X}_{i,t}$: The relative expected | 0.002* | 0.004*** | 0.002* | 0.003** |
| payoff of 60 | (0.001) | (0.001) | (0.001) | (0.001) |
| Indicator: 60 has the highest | 0.447*** | 0.440*** | 0.447*** | 0.440*** |
| expected payoff | (0.037) | (0.036) | (0.037) | (0.036) |
| Player fixed effects | Y | Y | Y | Y |
| Constant | 0.381*** | 0.385*** | 0.381*** | 0.385*** |
| | (0.034) | (0.034) | (0.034) | (0.034) |
| Obs | 11,810 | 11,810 | 11,810 | 11,810 |

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The table provides robustness checks to test if the identifed behavioral response is consistent with dart players having a hot-hand bias. The dependent variable is an indicator that equals 1 if the player shoots at the maximum-possible target of 60. $Hit_{t-1}$ is an indicator for whether the dart player's previous shot has hit the intended target, which is a proxy for having a hot hand. $\hat{X}_{i,t}$ is the estimated expected payoff of target 60, minus the estimated expected payoff of the next-highest target of 57, based on the current state of the world (including whether the player is in a hot state or not). All columns also include an indicator that equals 1 if $\hat{X}_{i,t} > 0$, i.e. if target 60 is calculated to be the optimal choice. Column 2 tests whether the behavioral response to a previous hit can be explained by being ahead or behind in the game, and includes the player's score minus the opponent's score. Column 3 tests whether the behavioral response to a previous hit can be explained by a desire to score a perfect 180 points in a turn. The model includes a dummy that equals 1 if the player is facing a third shot of a turn, and having hit 120 points with the previous two shots. Standard errors are clustered at the player level.

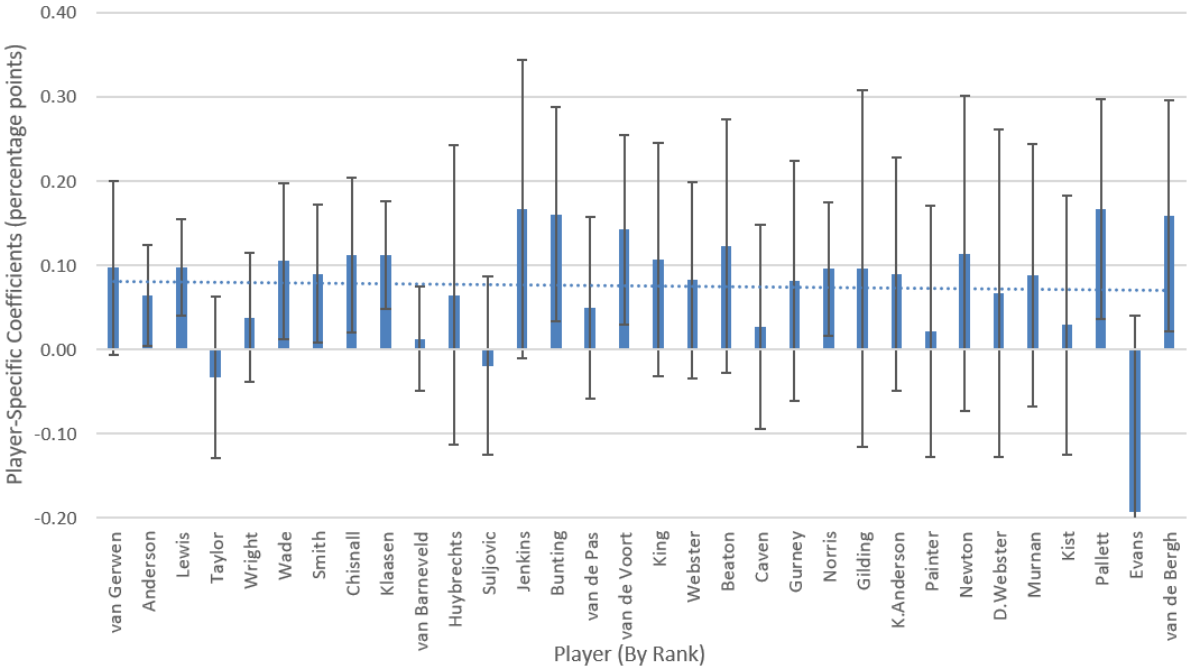



Figure 1.4.1: Player-Specific Hot-Hands

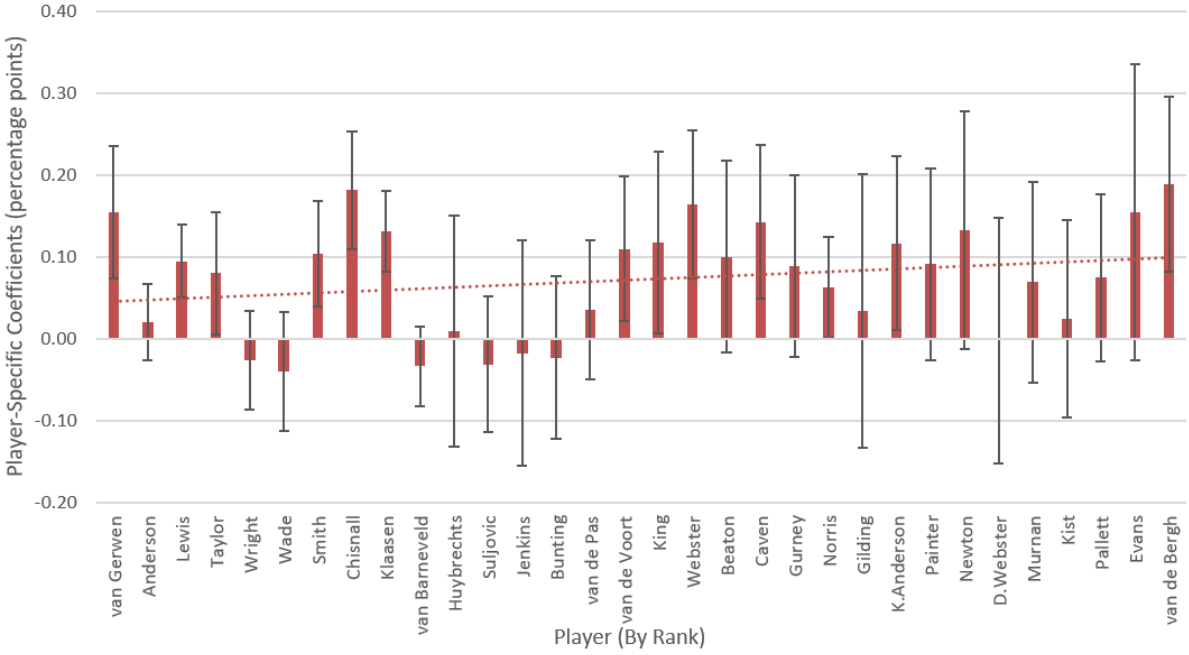



Figure 1.4.2: Player-Specific Hot-Hand Bias

and as a result have fewer observations. 29 out of 32 players have positive hot hand estimates, 11 of which are statistically significant at the 5% level. Only three players exhibit a negative hot hand, none significant at the 5% level. There appears to be no particular trend across the players' ranks.

Figure 1.4.2 shows player-specific estimates of hot-hand bias, using a specification similar to Column 3 of Table 1.3. The 95% confidence intervals are also shown. 25 players have positive hot-hand bias, 13 of which are significant at the 5% level. None of the estimates are significantly negative. The dotted line shows a slightly increasing linear trend, which is not significant ($p = 0.19$). If true, an increasing trend would imply that better players are less biased. Overall, the estimated hot-hand bias appears to be prevalent across most players.

## 1.5 Conclusion

This paper provides one of the first field evidence of hot-hand bias among professional athletes. Looking at the performance of professional dart players during a large international tournament, I find evidence of a large performance momentum – players are significantly more likely to be successful when their previous shot was a success. This is consistent with recent findings that significant hot hands exist in sports (Miller and Sanjurjo, 2014, 2015; Green and Zwiebel, 2017). Moreover, I find strong evidence that the majority of players misperceive their own hot hand – they are much too risk-taking when hot, consistent with hot-hand bias. This new evidence complements Green and Zwiebel (2017)'s findings that professional baseball pitchers tend to overreact to the *opponent'* most recent performances.

The hot-hand bias turns out to be costly in the sport of darts. Compared to a rational agent with correct perception of the hot hand, a biased player hits 0.57 fewer point per shot on average, or about 5.1 fewer points over the first nine shots. I estimate that this reduces the player's chance of winning the match by 1.4 percentage points, which translates to a $960 loss in prize earnings per match.

In the next chapter, I explore hot-hand bias in a medical setting. Specifically, I study physicians' treatment decisions during childbirth. I test whether a physician is more likely to choose C-section for a patient if the physician had previously performed a C-section successfully, which would be consistent with a belief in the hot hand of performing the surgery.

# Chapter 2

# Evidence of Hot-Hand Bias in Medical Decision-Making

## 2.1 Introduction

Estimates suggest that as much as $210 billion is wastefully spent on unnecessary medical treatments and services in the United States (McGinnis, Stuckhardt, Saunders, Smith, et al., 2013). In a survey conducted in 2009, 42% of primary-care physicians report that their patients receive too much care, in contrast to only 6% who said patients receive too little (Sirovich, Woloshin, and Schwartz, 2011). This paper focuses on a new concern stemming from the field of behavioral economics: the role of cognitive biases in physician decision-making. Specifically, this paper conducts an empirical investigation into the effects of hot-hand bias in the decision-making of obstetricians.

The "hot hand" refers to the notion that a person can enter a "hot" state in which her probability of success becomes higher than normal. For example, basketball players and fans believe that a player's chance of hitting a shot is greater following a hit than following a miss (Gilovich, Vallone, and Tversky, 1985). That is, a previous success is considered to be indicative of the player being in a hot state, and as a result people believe that the player has a higher chance of making the next shot. The "hot-hand bias" is when a person has exaggerated beliefs about the size of the hot hand. Studies have shown consistent evidence of hot-hand bias across a number of settings, for example in professional baseball where players have large hot hands (Green and Zwiebel, 2017) and even in gambling situations where hot hands clearly cannot exist because outcomes are known to be independent and identically distributed (Croson and Sundali, 2005; Guryan and Kearney, 2008; Suetens, Galbo-Jørgensen, and Tyran, 2015).

This paper uses micro-level data on childbirth to empirically test whether physicians have hot-hand bias when making treatment decisions. In childbirth, physicians must decide whether or not to perform a C-section on the patient. A physician who suffers from the hot-hand bias may, after having previously performed a C-section successfully, overestimate the likelihood of performing a successful C-section again. As a result, she may become more likely to choose C-section to the next patient than what is medically optimal for the patient.

I use administrative hospital discharge data for 1.3 million births in New York State over 2010-2015. First, I find no evidence that physicians have a hot hand when performing obstetrical procedures. In the absence of hot hand, physicians are still

2% more likely to perform a C-section after a previous successful C-section. The estimated effect is small but statistically significant ($p = 0.02$). My empirical model includes physician fixed effects, and a large set of patient conditions that proxy for when a C-section is likely to maximize patient welfare. Robustness checks provide additional evidence consistent with decision-makers having a hot-hand bias. The identified increase in C-section rate is not persistent, which is inconsistent with other explanations such as Bayesian learning process or malpractice fears. Generalizing the findings to the United States and assuming that the identified 2% increase in the C-section rate is unwarranted, the estimated health-care cost is $65 million per year.

This paper contributes to a growing literature on the role of cognitive biases in physician decision-making (e.g. Johnson and Goldstein, 2003; Blumenthal-Barby and Krieger, 2015; Khullar, Chokshi, Kocher, Reddy, Basu, Conway, and Rajkumar, 2015; Emanuel, Ubel, Kessler, Meyer, Muller, Navathe, Patel, Pearl, Rosenthal, Sacks, et al., 2016). It also contributes to the discussion of health-care overutilization in obstetrics, where health economics studies have explored the role of physician-induced demand (e.g. Gruber and Owings, 1996; Gruber, Kim, and Mayzlin, 1999; Johnson and Rehavi, 2016) and defensive medicine related to malpractice concerns (e.g. Dubay, Kaestner, and Waidmann, 1999; Currie and MacLeod, 2008; Shurtz, 2013). Finally, the empirical approach of this paper is motivated by studies on decision-making under the gambler's fallacy (Chen, Moskowitz, and Shue, 2016; Rabin and Vayanos, 2010; Suetens, Galbo-Jørgensen, and Tyran, 2015).

The paper is organized as follows. Section 2.2 describes the data, and Section 2.3 details the empirical specification. The results are presented in Section 2.4, and

Section 2.5 concludes.

## 2.2 Institutional Details and the Data

Childbirth can be performed vaginally or by a C-section. The C-section is a major abdominal surgery intended for the delivery of high-risk childbirths in which a vaginal delivery would put the baby or the mother at risk. Both the World Health Organization (WHO) and the American College of Obstetricians and Gynecologists (ACOG) recommend that a C-section should only be performed when medically necessary (Caughey, Cahill, Guise, Rouse, of Obstetricians, Gynecologists, et al., 2014; Betran, Torloni, Zhang, and Gülmezoglu, 2016). A C-section is $6,000 more expensive than vaginal births (Baicker, Buckles, and Chandra, 2006), and it is associated with an overall increase in poor outcomes for most pregnancies that are not high risk (Caughey, Cahill, Guise, Rouse, of Obstetricians, Gynecologists, et al., 2014).

In the U.S., the decision to perform a C-section is typically made by the physician, although mothers can also request the delivery method. There are several patient conditions that increase the risk of a C-section. C-sections are usually scheduled if the mother has a history of previous C-section, and about 90% end up receiving the surgery, even though many may benefit from first trying vaginal births after C-section (VBAC) (Caughey, Cahill, Guise, Rouse, of Obstetricians, Gynecologists, et al., 2014; Arnold and Flint, 2017). There are also important medical conditions that increase the chance of receiving a C-section, including a twin (or multiple) birth, problems with the placenta or the umbilical cord, obstructed labor, and breech

position. The C-section may either be scheduled ahead of time, or performed after an initial attempt of vaginal delivery.

There is a concern in the medical community that C-sections may be overused. The C-section rate has increased rapidly in the U.S. over the past 20 years, from 21% in 1995 to 32% in 2015 (Martin, Hamilton, Osterman, Driscoll, and Mathews, 2017). Studies suggest that overall birth outcomes do not improve beyond a C-section rate of about 20% (Gibbons, Belizán, Lauer, Betrán, Merialdi, Althabe, et al., 2010; Molina, Weiser, Lipsitz, Esquivel, Uribe-Leitz, Azad, Shah, Semrau, Berry, Gawande, et al., 2015). Perhaps more concerning is a large variation in C-section rates for low-risk births across hospitals, ranging from 7% to as high as 51% among major hospitals (Haelle, 2016). These numbers suggest a high level of discretion in performing C-sections.

The data set is provided by the Statewide Planning and Research Cooperative System (SPARCS) of New York. The data include all hospital admissions that occurred in the state of New York for the years 2010-2015, and contain a rich set of variables about the patient and the medical procedures performed during the admission.[1] I restrict the data to hospital admissions for childbirth, and then further restrict the sample to physicians who have performed at least one C-section and one vaginal delivery in the six-year period. The final analysis sample includes 1.3 million births in New York delivered by 3,725 physicians.[2]

[1] A complete list of variables is available at: https://www.health.ny.gov/statistics/sparcs/datadic.htm

[2] Prior to the New York data, I obtained a similar hospital discharge data from New Hampshire over 2010-2015. The data were provided by the Department of Health and Human Services of New Hampshire, and the analysis sample included 31,845 births delivered by 195 physicians. The results for New Hampshire is available in the Appendix.

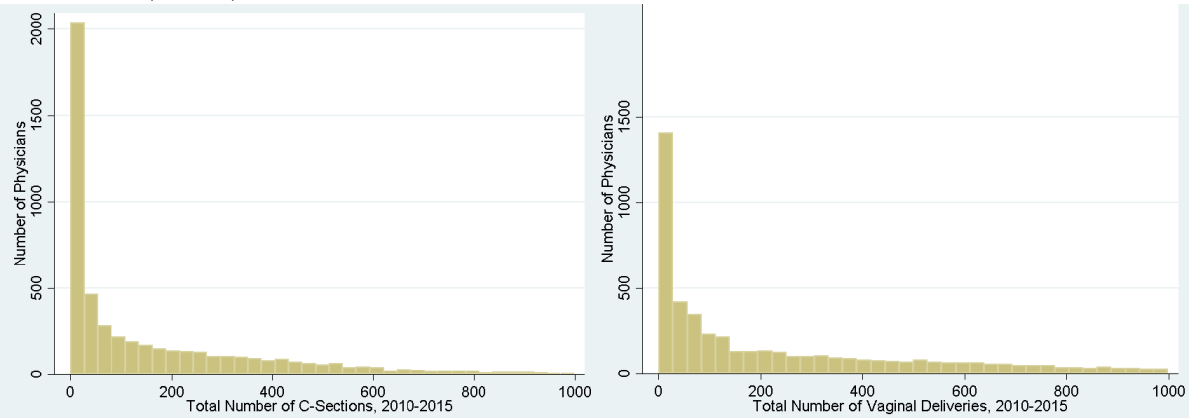
Table 2.1: Summary Statistics for Childbirth in New York State

| | Frequency (%) | |
|---|---|---|
| C-section rate | 34.2 | |
| | | |
| Any major complication | 12.7 | |
| *Maternal complication* | *6.1* | |
| *Fetal complication* | *7.1* | |
| | | |
| Patient Characteristics | Frequency (%) | % Receive C-Section |
| Age (years) | 29.4 | |
| History of previous C-section | 17.3 | 86.1 |
| Placenta previa | 0.8 | 71.7 |
| Disproportion | 0.2 | 87.3 |
| Breech position | 0.4 | 89.4 |
| Twin or multiple births | 2.2 | 66.4 |
| Pre-eclampsia | 4.8 | 53.6 |
| Hypertension | 9.4 | 48.1 |
| Diabetes | 1.0 | 53.5 |
| # Physicians | 3,725 | |
| # Births | 1,320,817 | |

Notes: The table shows summary statistics of the childbirth data obtained from all hospital admissions in New York State over 2010-2015. The sample is restricted to physicians who have performed at least one C-section and at least one vaginal delivery in the time period.

Table 2.1 shows summary statistics of the New York data set. The average C-section rate in New York is 34%. The rate is a little high because the sample excludes births delivered by family practitioners who do not perform C-section surgeries. The mean frequency of any major complication from births in New York is 12.7%. The maternal complication rate is 6.1%, and this includes hemorrhage, infections, and death. The fetal complication rate is 7.1%, and this includes hemorrhage, infections, fetal distress, trauma, and death. These rates are broadly in line with the complication rates in New Jersey (Currie and MacLeod, 2017).

Figure 2.2.1: Histograms Showing the Number of C-sections (Left) and Vaginal Deliveries (Right) by Physicians in New York State, 2010-2015




The average patient age in the New York sample is 29 years old. 17% of the patients have a history of previous C-section. Among these patients, 86% received a C-section, which is slightly lower than the national average of 91% in 2010 (Caughey, Cahill, Guise, Rouse, of Obstetricians, Gynecologists, et al., 2014). Other patient conditions, such as placenta previa and pre-eclampsia, are rarer, but are highly associated with receiving a C-section surgery.

Figure 2.2.1 plots two histograms showing the total number of C-sections (left) and vaginal deliveries (right) performed by New York State physicians in 2010-2015. More than 2,000 physicians have performed fewer than 36 C-sections (first bin) in my sample, and 55 physicians performed more than 1,000 C-sections (not shown). The distribution looks similar for vaginal deliveries. More than 1,400 physicians have fewer than 36 vaginal deliveries (first bin). 416 physicians have performed more than 1,000 vaginal deliveries.

## 2.3 Empirical Specification

I estimate the following OLS:

$$C_{i,t} = \alpha_i + \beta_1 S_{i,t-1} * C_{i,t-1} + \beta_2 S_{i,t-1} + \beta_3 C_{i,t-1} + \beta_4 X_{i,t} + \beta_5 X_{i,t-1} + \epsilon_{i,t}$$

$C_{i,t}$ is an indicator that equals 1 if physician $i$ performs a C-section on patient $t$. $S_{i,t-1}$ is an indicator that equals 1 if physician $i$'s previous patient $t$'s childbirth was successfully administered without any major maternal or fetal complication (see Section 2.4.1). The regression includes physician fixed effects, $\alpha_i$, and controls for patients $t$ and $t-1$, denoted by vectors $X_{i,t}$ and $X_{i,t-1}$, respectively. The standard errors are clustered at the physician level.

$X_{i,t}$ contains the following key patient characteristics: patient $t$'s age; an indicator for 40 years or older; dummies for race; a history of previous C-section; placenta previa; disproportion; breech position; multiple birth; pre-eclampsia; hypertension; diabetes; day-of-week of the procedure; month; year; and hospital fixed effects.

The coefficient of interest is $\beta_1$. It is interpreted as the percentage point change in the C-section rate following a successful C-section, conditional on the set of controls included in the model. $\beta_1 = 0$ would be consistent with a physician who decides the treatment solely based on the medical conditions of the patient, as recommended by standard medical guidelines (Caughey, Cahill, Guise, Rouse, of Obstetricians, Gynecologists, et al., 2014; Betran, Torloni, Zhang, and Gülmezoglu, 2016). $\beta_1 > 0$ would be consistent with hot-hand behavior. That is, physicians increase the C-section rate after a successful C-section because of a belief that a subsequent C-

section is more likely to be successful. This behavioral response could be optimal if physicians actually have a hot hand; otherwise, it would imply a bias.

Because I include physician fixed effects, the identification comes from within-physician variations. Broadly speaking, the identifying assumption is that conditional on a set of controls I use, the patients treated after a successful C-section are similar to those who are not.[3] More precisely, the identifying assumption is that physicians believe that patients are drawn from the same distribution, independent of whether the previous delivery was a successful C-section.

In Section 2.4.4 I discuss possible ways in which the identifying assumption may be violated. For example, physicians may schedule a C-section after an "easy" C-section. This would imply that patients in the treatment group will have higher C-section risk than the control group, and as a result $\beta_1$ would be biased upwards. One approach to address this problem is to consider only weekends and holidays, for which deliveries are generally not scheduled. Another approach is to drop from the sample multiple C-sections performed by a physician in a day. The results from both approaches suggest that endogenous scheduling of C-sections is unlikely to be driving the main results.

[3]The latter category includes previous unsuccessful C-section, and previous vaginal deliveries (successful or unsuccessful).

## 2.4 Results

### 2.4.1 Test for the Hot-Hand

The medical guidelines recommend that treatment decisions should depend on the conditions of the patient alone (Caughey, Cahill, Guise, Rouse, of Obstetricians, Gynecologists, et al., 2014; Betran, Torloni, Zhang, and Gülmezoglu, 2016). However, if physicians have a hot hand, then it may be in the patients' best interest for treatment decisions to be influenced by whether the physician has a hot hand. More specifically, if a physician is more likely to succeed in performing a C-section after a successful C-section, or if she is less likely to succeed in performing a vaginal delivery after a successful C-section, it might be optimal for her to temporarily increase the C-section rate in response to this hot hand.

I find no evidence of a hot hand in performing obstetrical procedures. Table 2.2 shows estimates from two regressions, where the dependent variable is an indicator for a successful procedure. Column 1 focuses on C-section deliveries and shows that a prior successful C-section is not positively associated with the probability of success of a subsequent C-section. Similarly, Column 2 focuses on vaginal deliveries, and I find that a prior success performing a C-section is not significantly associated with the probability of success of a subsequent vaginal delivery.

### 2.4.2 Test for the Hot-Hand Bias

This section explores whether physicians' decision to perform a C-section is associated with whether the previous delivery was a successful C-section.

Table 2.2: Test for the Hot-Hand (Childbirth)

| | (1) | (2) |
|---|---|---|
| Dep Var = Success | C-section | Vaginal Delivery |
| Previous delivery was | -0.001 | -0.001 |
| C-section * success | (0.004) | (0.002) |
| | | |
| Previous delivery was a success | 0.002 | 0.005*** |
| | (0.002) | (0.002) |
| Previous delivery was a C-section | 0.001 | 0.000 |
| | (0.004) | (0.002) |
| | | |
| Controls | Y | Y |
| Mean of Dep Var | 0.85 | 0.88 |
| # Births | 451,132 | 869,685 |
| # Physicians | 3,722 | 3,667 |

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The table tests whether physicians exhibit a hot hand when performing obstetrical procedures. The dependent variable is an indicator that equals 1 if the procedure (C-section in Column 1; vaginal delivery in Column 2) is completed successfully, without major maternal or fetal complications. The explanatory variable of interest is the interaction term, which equals 1 if the physician's previous delivery was a successful C-section. Standard errors are clustered at the player level.

In Table 2.3, Column 1 shows regression estimates using the full analysis sample, and controlling for physician fixed effects, hospital fixed effects, as well as dummies to control for the day-of-week, month, and year of the procedure. After a successful C-section, the C-section rate increases by 1.1 ppts. A previous successful vaginal delivery is negatively associated with the C-section rate, but it is not significant at the 5% level. The C-section rate appears to be positively associated with having performed a C-section previously.

In Column 2, I include patient conditions for the current ($t$) and previous ($t-1$) patients. The estimated coefficient of interest is smaller: after a successful C-section, the C-section rate increases by 0.7 ppts ($p = 0.02$). This corresponds to a 2% increase from the mean C-section rate of 34%. Whether the previous delivery was successful, or whether it was a C-section, are both not significantly associated with the subsequent C-section rate.

If the estimated effect is indeed because of a belief in the hot hand, the effect should primarily be driven by physicians whose previous delivery was recent. For example, a hot hand belief is a plausible explanation if the physician has performed a successful C-section yesterday, but less plausible if the C-section was more than, say, two weeks ago.

Column 3 restricts the sample to childbirths in which the physician's previous delivery was less than two days ago. The sample size drops to about half: 738,431 births and 3,230 physicians. The point estimate for the interaction term is slightly larger: after a recent successful C-section (less than two days ago), the C-section rate increases by 0.9 ppt (3%) ($p = 0.012$).

Table 2.3: Test for the Hot-Hand Bias (Childbirth)

| | (1) | (2) | (3) |
|---|---|---|---|
| Dep Var = C-section | Full Sample | Full Sample | Subsample |
| Previous delivery was | 0.011*** | 0.007** | 0.009** |
| C-section * success | (0.003) | (0.003) | (0.003) |
| | | | |
| Previous delivery was a success | -0.004* | -0.002 | -0.004* |
| | (0.002) | (0.002) | (0.002) |
| Previous delivery was a C-section | 0.010*** | 0.004 | 0.008** |
| | (0.003) | (0.003) | (0.004) |
| | | | |
| Physician FE, Hospital FE | Y | Y | Y |
| Day-of-week FE, Month FE, Year FE | Y | Y | Y |
| Patient characteristics $(X_{i,t}, X_{i,t-1})$ | | Y | Y |
| | | | |
| Constant | 0.254*** | 0.117*** | 0.105*** |
| | (0.003) | (0.004) | (0.005) |
| Mean of Dep Var | 0.34 | 0.34 | 0.32 |
| # Physicians | 3,725 | 3,725 | 3,230 |
| # Births | 1,320,817 | 1,320,817 | 738,431 |

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The table tests whether the decision to perform C-section is related to having previously performed a successful C-section. The dependent variable is an indicator that equals 1 if the physician performs C-section on the current patient. The explanatory variable of interest is the interaction term, which equals 1 if the physician's previous delivery was a successful C-section. This is a proxy for having a hot hand. Column 1 does not include patient characteristics. Column 2 is the preferred model, with patient characteristics included as controls. Column 3 restricts the sample to deliveries where the physician's previous delivery was less than two days ago. Standard errors are clustered at the player level.

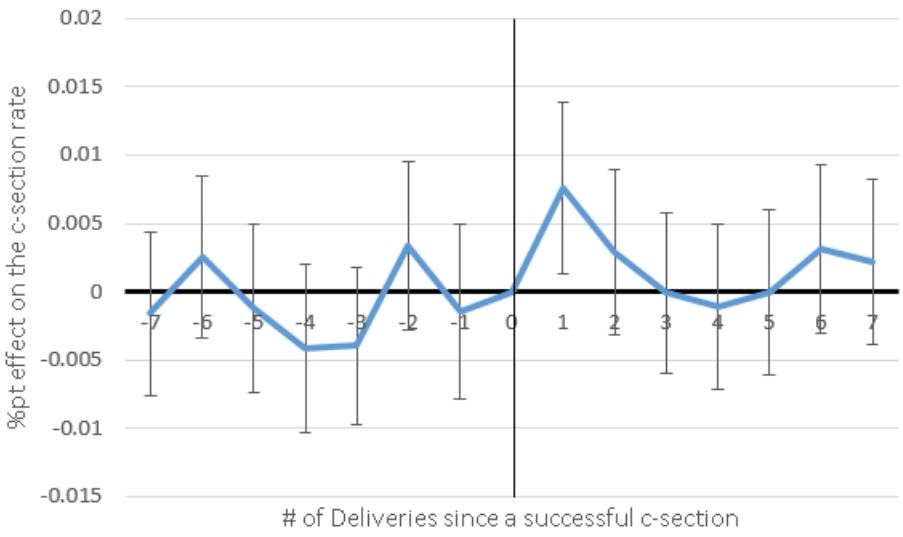



Figure 2.4.1: The Effect of Previous Successful Surgery on C-section Rate (New York)

Next, I add 6 additional lags and 7 leads of the three main explanatory variables, as well as controls for the 13 patients, to the main regression shown in Column 2. Figure 2.4.1 plots the 14 estimated coefficients for the interaction term, with 95% confidence intervals. The $x$-axis shows the number of deliveries since a successful C-section, and the $y$-axis shows the percentage point effect on the C-section rate. Negative $x$ correspond to coefficients for leads, and positive $x$ correspond to coefficients for lags. The point estimates for leads are a little noisy, but none are statistically significant at the 5% level. Immediately after a successful C-section, the C-section rate increases by 0.8 ppt ($p = 0.018$), but the effect does not persist and diminishes quickly for longer lags.

Table 2.4: Robustness Checks (Childbirth)

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dep Var = C-section | Ever CS | Never CS | Weekends | Weekdays |
| Previous delivery was | -0.008 | 0.008*** | 0.010 | 0.006* |
| C-section * success | (0.006) | (0.003) | (0.006) | (0.003) |
| Controls | Y | Y | Y | Y |
| $p$-value for difference in coefficients | 0.009 | | 0.291 | |
| Mean of Dep Var | 0.86 | 0.23 | 0.26 | 0.37 |
| # Physicians | 3,282 | 3,693 | 3,400 | 3,701 |
| # Births | 228,622 | 1,092,195 | 286,481 | 1,034,336 |

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The table conducts two robustness checks. The dependent variable is an indicator that equals 1 if the physician performs C-section on the current patient. The explanatory variable of interest is the interaction term, which equals 1 if the physician's previous delivery was a successful C-section. Columns 1 and 2 test the prediction that the effect diminishes for patients who have a history of previous C-section ("Ever CS"). Columns 3 and 4 test the prediction that the hot-hand bias is larger during weekends than during weekdays. Standard errors are clustered at the player level.

### 2.4.3 Robustness Checks

In this section I conduct several robustness checks. First, I restrict the analysis to a subsample of mothers with a history of having had a previous C-section. The idea is that the hot state of the physician should have weak or no influence on the method of delivery for these patients. A majority of these patients schedule a C-section ahead of time, and those who do not tend to have a preference for a vaginal birth after a C-section. In both cases, the physician's previous surgical success is unlikely to have an effect on the treatment decision.

In Column 1 of Table 2.4, I find a non-significant negative effect among mothers with a history of previous C-section ("Ever CS"). By contrast, the estimated effect

is positive and significant in the subsample of mothers who never had a C-section before (Column 2). The two coefficients are statistically different from each other ($p = 0.009$).

Next, I test the prediction that hot-hand bias is likely to be larger during weekends than during weekdays. This is because the bias cannot affect treatment decisions for C-sections that are planned ahead of time. C-sections are rarely planned during weekends, and so the estimated effect may be larger than during weekdays where there is a mix of planned and unplanned deliveries. In Column 3, the point estimate is slightly larger for weekends, but the difference is not statistically significant.

### 2.4.4 Discussion of Mechanisms

To summarize, I find that a previous successful C-section is associated with a 2% increase in the subsequent C-section rate. The effect is not persistent beyond the immediate delivery after a successful C-section. It is consistent with a hot hand belief, that is, a physician is more likely to perform a C-section right after a successful C-section because of a belief that a subsequent C-section is more likely to succeed. I also find some weak but suggestive indication that a successful vaginal delivery is negatively associated with the physician's subsequent C-section rate. If true, this would be consistent with a hot hand belief in performing vaginal deliveries.

Since treatment decisions normatively should depend on the conditions of the patient alone, standard explanations are generally less consistent with the results. For example, I find no empirical evidence that physicians exhibit a hot hand in performing obstetrical procedures.

The fact that the estimated hike in C-section rate is not persistent at all helps rule out several alternative mechanisms. For example, a reasonable learning process – Bayesian or otherwise – would have a more persistent effect on the subsequent C-section rate. Likewise, salience effects or a defensive response to malpractice fears would also generate more persistent effects. For example, Shurtz (2013) find a 4% discontinuous jump in the C-section rate that persists after the physician has made a medical error.

Another possible mechanism is a belief in luck reversals (Rabin and Vayanos, 2010; Chen, Moskowitz, and Shue, 2016). After a successful delivery, the physician might believe that the next childbirth is less likely to be successful, and as a result perform C-sections as a defensive measure. This could explain why the C-section rate increases following a successful C-section, but it would also suggest that the C-section rate should increase following a successful vaginal delivery. In fact, I find weak evidence of the opposite effect.

Another possibility is that physicians endogenously schedule C-sections together in a way that would create an upward bias in the estimate. For example, physicians may have a tendency to schedule a C-section after an "easy" C-section. To address this possibility, I have identified 129,398 C-sections that correspond to multiple C-sections performed by a physician in a day. These are likely candidates for endogenous scheduling. When I re-estimate the baseline regression without these potentially endogenously-scheduled C-sections, the point estimate for the interaction term remains intact at 0.7 ppts ($p = 0.014$).

## 2.5 Conclusion

This paper finds a small but significant increase in the C-section rate if the physician has previously performed a C-section successfully. The model controls for physician fixed effects and a large set of observable medical conditions of the patient to proxy for when a C-section is likely to maximize patient's welfare. The evidence is most consistent with hot-hand bias. The identified hike in the C-section rate is not persistent, which is inconsistent with alternative explanations such as malpractice fears and Bayesian learning process that predict more persistent effects.

The economic consequences of hot-hand bias in obstetrics are large. Suppose that the identified 2% increase in the C-section rate is entirely unnecessary.[4] Generalizing to the US, this would create 8,000 unnecessary C-sections each year. According to the New York hospital data, a C-section is about $7,000 more expensive than natural birth, which sums up to $55 million in unnecessary health-care costs. In addition, mothers who receive unnecessary C-sections are more likely to receive C-section again in the future. I estimate this additional cost to be $10 million, bringing the total estimate to $65 million in unnecessary health-care costs per year in the US. This estimate is conservative, because it excludes disutility from receiving a major abdominal surgery and potential long-term health risks associated with C-sections for low-risk births.

[4]The assumption is that decision errors only occur in one direction where the mothers who do not medically require a C-section can receive unnecessary C-section, but mothers who medically require a C-section will always receive a C-section. This assumption is based on the evidence that C-sections are only medically necessary for about 20% of the patients (Gibbons, Belizán, Lauer, Betrán, Merialdi, Althabe, et al., 2010; Molina, Weiser, Lipsitz, Esquivel, Uribe-Leitz, Azad, Shah, Semrau, Berry, Gawande, et al., 2015), much lower than the current rate of 32% in the U.S. (Martin, Hamilton, Osterman, Driscoll, and Mathews, 2017).

It is challenging to think about policy interventions that can mitigate the hot-hand bias. Will telling physicians about the bias reduce it? Will showing electronic reminders of the medical guidelines for when to perform C-sections effective in mitigating the hot-hand bias? One possible approach is to conduct an incentivized lab experiment with resident physicians. The experiment could simulate the medical decision-making process by presenting the medical students with patient records, asking them to make treatment decisions, and showing them the outcome. I could then implement various interventions in the lab to test if they reduce the hot-hand bias. Given that I find a small effect in the field, it would be important to check if there will be enough power to identify the effects of interventions in a lab setting.

# Chapter 3

# Sleep and Health: Evidence from Daylight Saving Time (with Nicolas Ziebarth)

## 3.1 Introduction

Sleep deprivation is becoming a major public health concern in many developed countries around the world. The global sleep-aid market is growing rapidly with an estimated size of $80 billion in 2020 (Persistent Market Research, 2015). The United States alone counts 40 million sleeping pill prescriptions per year and about 2800 "sleep labs" exist (CDC, 2013; DiSalvo, 2015). Hillman, Murphy, Antic, and Pezzullo (2006) estimate the economic costs of sleeplessness at almost one percent of GDP.

There is rich medical literature on the relationship between sleep and health. Numerous studies have documented positive associations between sleep deprivation, poor health, and decreased cognitive ability, but it remains unclear whether this link represents a causal relationship (Moore, Adler, Williams, and Jackson, 2002; Taheri, Lin, Austin, Young, and Mignot, 2004; Mullington, Haack, Toth, Serrador, and Meier-Ewert, 2009; Killgore, 2010). Banks and Dinges (2007) provide a comprehensive review of the behavioral and physiological effects of inadequate sleep, including experimental evidence with healthy adult laboratory participants. They conclude that restricting sleep below an individual's optimal could cause a range of neurobehavioral deficits.

Fewer studies have investigated the role of sleep in the economics literature. Biddle and Hamermesh (1990) show that increased labor market activities reduce sleep duration. Hamermesh, Myers, and Pocock (2008) exploit television schedules and time use data to demonstrate how time zones affect market work and sleep in the US. Giuntella and Mazzonna (2017) also exploit US time zones to show in a geographic Regression Discontinuity Design that sleep deprivation can lead to poor health and obesity. Moreover, Gibson and Shrader (2018) identify positive wage returns to sleep. Billari, Giuntella, and Stella (2017) exploit the rollout of high-speed internet access in Germany and show that DSL access reduces both sleep duration and sleep satisfaction.

This paper investigates the short-term causal effect of getting additional sleep on health. We exploit the quasi-experimental nature of a regulation that has been affecting the sleep pattern of more than one billion people in 70 countries around the

globe: Daylight Saving Time (DST). It is the practice of setting clocks forward by one hour in spring and backward by one hour in fall. The original DST rationale was to save energy. Today, all countries in the European Union, the great majority of the US states and Canadian provinces, as well as 40 other countries such as Mexico, Chile, Israel, and Iran observe DST.

Our identification strategy focuses on the time shift in the fall when the clocks "fall back" and exogenously add one additional hour at night. The main idea is that this extension at night induces people to sleep more. Indeed, using a large US survey, we find that people report sleeping significantly more following the time shift. Moreover, we find a significant reduction in the share of people who report having unintentionally fallen asleep during the day. We then combine large survey and administrative hospital data to identify the health benefits of getting more sleep at the population level.

We use two large datasets that complement each other: (a) The US Behavioral Risk Factor Surveillance System (BRFSS), which records sleep and self-reported health and allows us to study mild and subjective health effects; and (b) The German Hospital Census, which records all hospitalizations and allows us to study objective health effects. Both datasets together provide evidence from the most populous American and European country over a decade. Both datasets carry large numbers of observations – 3.4 million interviews from the U.S. and 160 million hospital admissions from Germany. The large sample sizes are crucial to control for seasonal and weekday confounders while maintaining enough statistical power to precisely identify health effects at a daily level.

Our findings show consistent evidence that both subjective and objective health measures significantly improve for about four days after getting more sleep. The BRFSS data show that the share of US citizens reporting "excellent health" increases from 19 to 20% between days 1 to 4 after the time shift. In addition, hospital admissions decrease significantly, and this effect also persists for about four days. For example, hospitalizations due to cardiovascular diseases decrease by ten admissions per day per one million population. We then discuss alternative mechanisms through which the DST transition might affect health, and show that these are unlikely to drive our findings.

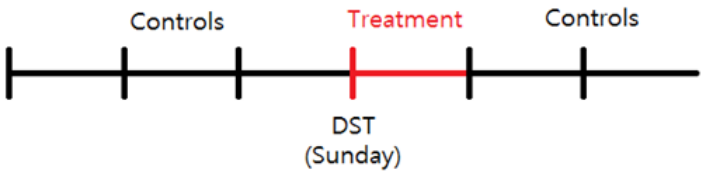
The next section briefly describes the data. Section 3.3 outlines the empirical methodology. Section 3.4 presents and discusses the findings and Section 3.5 concludes.

## 3.2 Datasets

### 3.2.1 The US Behavioral Risk Factor Surveillance System (BRFSS)

The first dataset measures sleep duration and subjective health effects in the general population. The Behavioral Risk Factor Surveillance System (BRFSS) is a large, annual telephone survey of US adults aged 18 and above, administered by the Centers for Disease Control and Prevention (CDC). The survey began in 1984 with fifteen participating states; by 1996, all 51 US states participated in the survey. It covers an extensive set of self-reported health and also sleep. It is, by design, representative

Figure 3.2.1: Sample Selection of Main Models–Extracting 6 Weeks around DST Change




of state populations. We focus on the period from 2001 to 2010 which includes more than 3.4 million survey responses in total. However, as illustrated by Figure 3.2.1, our main sample extracts six weeks around the time shift and counts 421,101 observations.

The BRFSS includes demographics such as age, sex, race, and marital status, as well as education and employment status.

### Construction of Main Dependent Variables

First, we use the standard self-assessed health (SAH) question: "*Would you say that in general your health is ___?*" The majority of respondents report their general health to be either very good (32%) or good (30%), and about 19% report excellent general health. Less than 6% of the population report poor general health. From this, we construct two binary dependent variables: (a) *Excellent health*, and (b) *fair or poor health*.

Second, we use two sleep measures. Responses to the following question are integers between 0 and 24: "*On average, how many hours of sleep do you get in a 24-hour period? Think about the time you actually spend sleeping or napping, not just the amount of sleep you think you should get.*" We interpret the answers as a

good measure of actual sleep duration. It is worth noting that the question does not explicitly ask for the duration of sleep *last night*, but instead the responses will reflect average sleep in the recent past. Hence, our estimate on sleep duration is likely downward biased and a lower bound. Thus, it provide a conservative test whether people sleep more when clocks are set back in fall.

We also use responses to the following question to measure tiredness during the day: "*During the past 30 days, for about how many days did you find yourself unintentionally falling asleep during the day?*" We convert the responses into a binary variable indicating the share of people who unintentionally fell asleep. On average, 35% of the US population report unintentionally falling asleep.

Finally, in robustness checks and falsification testes, we use information on whether respondents received a flu shot in the past calendar year, and whether they exercise.

### Daylight Saving Time in the US

In the United States, DST ends on the first Sunday in November. The time change occurs at 2am, where the clocks are set back to 1am, effectively extending the night by one hour. Table 3.1 shows the dates of the time shift from 2000-2010. DST is observed by most states in the US. As of 2018, the states that do not observe are Arizona, Hawaii, and overseas territories. Indiana only began to observe DST in 2006. Our empirical strategy only uses states that observe DST.

Table 3.1: Date of Fall DST Transition in Germany and the US

| Year | DST Fall US | DST Fall Germany |
|---|---|---|
| 2000 | 10/29/2000 | 10/29/2000 |
| 2001 | 10/28/2001 | 10/28/2001 |
| 2002 | 10/27/2002 | 10/27/2002 |
| 2003 | 10/26/2003 | 10/26/2003 |
| 2004 | 10/31/2004 | 10/31/2004 |
| 2005 | 10/30/2005 | 10/30/2005 |
| 2006 | 10/29/2006 | 10/29/2006 |
| 2007 | 11/4/2007 | 10/28/2007 |
| 2008 | 11/2/2008 | 10/26/2008 |
| 2009 | 11/1/2009 | 10/25/2009 |
| 2010 | 11/7/2010 | 10/31/2010 |

### 3.2.2 German Hospital Admissions Census

The second dataset provides objective health measures. The dataset comprises all German hospital admissions from 2000 to 2008. By law, German hospitals are required to submit depersonalized information on every single hospital admission. The 16 German states collect these information and the German Federal Statistical Office provides restricted data access for researchers. Germany has about 82 million inhabitants and about 17 million hospital admission per year. To obtain the working dataset, we aggregate the admission-level data on the daily county level and then normalize admissions per 100,000 population.

The data include information on age and gender, the day of admission, the county of residence as well as the diagnosis in form of the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) code.

As with BRFSS, our working dataset focuses on the six weeks centered around the time shift (Figure 3.2.1). This main sample has 336,604 county-day observations

over 9 years. We leave the data at the county-level and do not further aggregate up to the national level for a few reasons. First, this allows us to stratify the effects by county characteristics and weather and pollution conditions. Another reason is that we lose statistical power when aggregating up to a time series at the national level.

### Construction of Main Dependent Variables

Using the information on primary diagnosis, we generate the following dependent variables: *All cause admission rate*. On a given day, we observe 59.77 hospital admissions per 100,000 population. However, the rate varies substantially and the standard deviation is 25.73.

By extracting the ICD-10 codes I00-I99—diseases of the circulatory system—we generate *Cardiovascular admission rate*. This is the single most important subgroup of admissions—9.53 admissions per 100,000 population account for 16% of all admissions. Extracting the codes I20 and I21, the *Heart attack rate* is 1.59 admissions per 100,000 population.

Finally, we generate the *injury rate* (V01-X59) as well as the *respiratory* (J00-J99), *metabolic* (E00-E90), *neoplastic* (C00-D48), and *infectious admission rate* (A00-B99). We also test for changes in *suicide attempts* (T14) and *drug overdosing* (T40) per 1 million population.

### Daylight Saving Time in Germany

In Germany, DST ends on the last Sunday of October in all German states (Table 3.1). The time change occurs on 3am where the clocks are set back to 2am. Again,

for the main analysis, we restrict our sample to six weeks around the time change (Figure 3.2.1).

## 3.3 Empirical Specification

Our identification strategy relies on sleep extensions created by DST transitions in the fall. These occur on different dates each year. Our large datasets allow us to comprehensively control for seasonal confounders, weekday effects, and yet to precisely estimate health effects. Our preferred empirical specification identifies the effects at the daily level. We also estimate models at the weekly level to capture medium-term and potential intertemporal substitution effects.

### 3.3.1 Main Specification

Our preferred specification employs daily dummies around the DST time shift in the fall:

$$y_{id} = \beta_0 + \beta_1 DST_{id} + X'_{id}\gamma + Vacation_d + \phi_m * \delta_t + DOW * \phi_m + t + t^2 + \mu_s + \epsilon_{id} \quad (3.3.1)$$

Where $y_{id}$ is the health outcome variable using the German Hospital Census (BRFSS), for county (individual) $i$ on day $d$. $DST$ is a vector containing fifteen daily dummies around the DST time shift.

Equation (3.3.1) includes controls that net out seasonal and weekday confounders. These are crucial when using high-frequency data within the DST context. For

example, hospital admissions decrease on Sundays and also on national holidays (Witte, Grobbee, Bots, and Hoes, 2005). $Vacation_d$ controls for public holidays and the Halloween.

Due to the relevance of day-of-the-week (DOW) effects, we additionally interact DOW with month fixed effects ($DOW * \phi_m$). This is important as Sundays in November may be systematically different from Sundays in September. For example, in our data, relative to Sundays, hospital admissions almost double on Mondays and this effect varies over the months of a year. Because DST transitions are always on Sundays, it is crucial to net out DOW effects by month of the year.

Our model also includes month-year fixed effects ($\phi_m * \delta_t$) and linear and quadratic time trends ($t + t^2$). However, the findings are robust to replacing month-year fixed effects with separate month and year fixed effects and omitting time trends. In addition, Equation (3.3.1) corrects for county-level or individual-level socio-demographics ($X'_{id}\gamma$) and persistent differences across states or counties ($\mu_s$).

Because it is unlikely that admission rates are either independent over time or across space, we correct the standard errors, $\epsilon_{id}$, by applying two-way clustering across counties and over time (Cameron, Gelbach, and Miller, 2011). When using the independently drawn and representative observations of the BRFSS, we cluster standard errors at the date level. All BRFSS regressions are probability weighted.

### 3.3.2 Identification

The key idea of our identification strategy is that the running variable is time; and that the DST transition generates the treatment. DST transitions are arguably ex-

ogenous to individuals because humans cannot influence time. Our main specification de-trends the outcome variables using DOW-month and month-year fixed effects, in addition to socio-demographic controls. We also disentangle weekday and seasonal effects from specific events such as vacation days or national holidays. The richness of our data still allows us to obtain precise estimates at the daily level. However, we also compare the day-to-day short-term effect of the change in time to the net effect on a weekly basis. Moreover, in effect heterogeneity specifications that test for behavioral mechanisms, we stratify the results by ambient climatic conditions such as temperatures, hours of sunshine, and pollution.

**Sample Selection and Definition of Treatment and Control Groups**

As illustrated in Figure 3.2.1, we restrict our main sample to three weeks before and three weeks after the time shift. However, the results are robust to including all 52 weeks of the year. The findings are also robust to assigning all three post-transition weeks to the "treatment group." Doing this yields results that are similar to a standard Regression Discontinuity design where the post-treatment outcomes are compared to that of the pre-treatment, conditional on all covariates, see for example Doleac and Sanders (2015).

## 3.4 Results

### 3.4.1 The Effect of Time Shift on Sleep

First, we use BRFSS measures on sleep to provide first-stage evidence that the fall DST time shift increases the average sleep duration in the population. In the US, on the first Sunday of November, the clocks are set back by one hour from 2am to 1am, effectively extending night time by an hour. (In Germany, the clocks fall back from 3am to 2am on the last Sunday of October; Table 3.1.)

Table 3.2 shows the results when we estimate Equation (3.3.1) using the BRFSS sleep measures as outcome variables. The first two columns use self-reported hours of sleep as the dependent variable; the last two columns use unintentionally fell asleep as dependent variable. As discussed in Section 3.2.1, our measures capture sleep in the *recent past*, not just from last night. The estimates are therefore likely downward-biased; for us, they provide a crude affirmative test that people do sleep more when clocks fall back.

According to column (1), on average, people sleep an additional 0.27 hours (or about 16 minutes). This estimate is statistically significant at the 1% level. The regression includes the same set of controls as our preferred model in Equation (3.3.1), comprehensively netting out seasonal effects. Note that this is an average effect across the entire population. It is likely driven by the sleep deprived. In fact, a 16-minute increase in sleep across the entire population is consistent with a quarter of the population sleeping one hour more.

Column (2) uses a model that measures the effect at the weekly level. Here we

Table 3.2: Effects of Fall DST Transition on Sleep

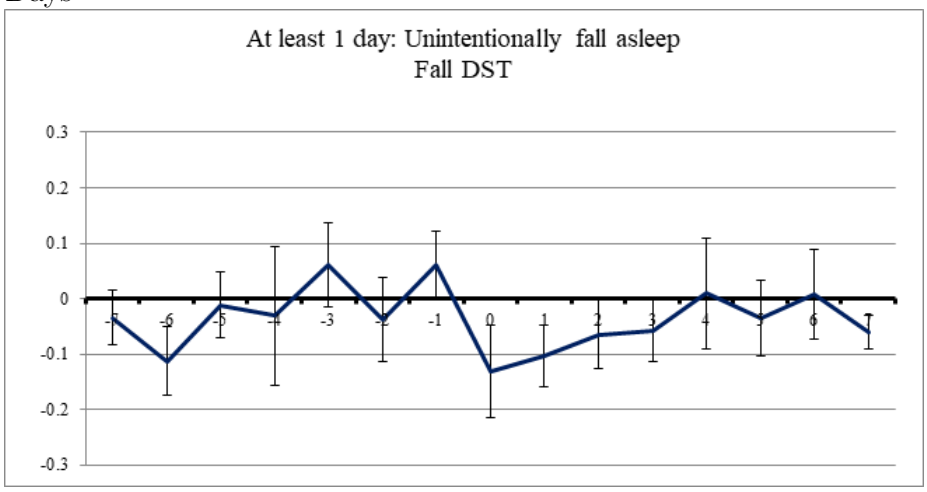
| | (1) Hours of Sleep | (2) Hours of Sleep | (3) Unintentionally falling asleep | (4) Unintentionally falling asleep |
|---|---|---|---|---|
| Day of Transition | 0.265*** | | -0.061 | |
| | (0.079) | | (0.054) | |
| Week of Transition | | 0.182*** | | -0.044** |
| | | (0.069) | | (0.022) |
| Controls | X | X | X | X |
| *Dep. var. mean* | 7.07 | 7.07 | 0.35 | 0.35 |
| $R^2$ | 0.06 | 0.06 | 0.07 | 0.07 |
| Observations | 10,833 | 10,833 | 10,833 | 10,833 |

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the date level. Regressions are probability-weighted. Day of Transition is an indicator variable that equals 1 if the interview is on the day of DST transition in the fall. Week of Transition is an indicator that equals 1 if the interview is on the Sunday of DST transition or one of the following 6 days. Controls include state fixed effects, indicator for the Halloween, day-of-week X month fixed effects, month x year fixed effects, linear and quadratic time trends, and socioeconomic covariates. In 2009, six states (Georgia, Hawaii, Illinois, Louisiana, Minnesota, and Wyoming) began to include questions about sleep in the BRFSS; this expanded to nine states in 2010 (Arkansas, Connecticut, Delaware, District of Columbia, Hawaii, Minnesota, Missouri, Nevada, and Oregon). The column headers describe the dependent variables used in each column; columns (1) and (2) have values between 0 and 24; columns (3) and (4) use binary measures. Each column is one model as in Equation (3.3.1).

include a dummy variable that equals 1 for the entire the week of the DST transition. Again, we find that people sleep a statistically highly significant 0.18 hours (or 11 minutes, 2.5%) more per night (for seven nights) on the week of DST transition when the night hours are extended.

We find corroborating evidence when we turn to self-reported measures of tiredness in columns (3) and (4). The estimated daily effect in column (3) is negative but imprecisely estimated, which may be due to the noisy nature of this survey question.

Figure 3.4.1: Effects of Fall DST Transition on Unintentionally Falling Asleep in Past 30 Days




In column (4), where we estimate the weekly model, we find that people are 4.4 percentage points (ppts) or 12.6% less likely to report having fallen asleep in the week of time shift. This estimate is statistically significant at the 5% level.

Figure 3.4.1 plots the daily dummies of the vector $DST_{id}$ in Equation (3.3.1) using fell unintentionally asleep as outcome measure. Figure 3.4.1 is an event study-type graph and plots estimates for -7, -6,. . . ,0,. . . , 6, 7 days relative to the time shift. Note that this is not a simple descriptive graph but compares the effect in the treatment group relative to the control group, after having netted out of seasonal and weekday confounders (Equation (3.3.1)).

In Figure 3.4.1, despite the noisy nature of the self-reported measure, one observes a distinct four-day decrease in tiredness following the DST transition. As we discuss below, we find very similar four-day health improvements using different health measures from the US and Germany, such as self-reported health or hospi-

tal admissions. We interpret this consistent pattern as reinforcing evidence for the credibility of our identification strategy.

### 3.4.2 The Effect of Time Shift on Hospital Admissions

Next, we study whether hospital admissions vary significantly as a result of the time change. Table 3.2 shows that people report having significantly longer sleep on the day and in the week of the fall transition when clocks are set back in the middle of the night. We expect the effects to be particularly concentrated among the sleep deprived—studies show that about ten percent of the population are permanently sleep deprived (e.g. Knutson, Van Cauter, Rathouz, DeLeire, and Lauderdale, 2010).

Table 3.3 shows weekly admission estimates by disease groups for Germany. Each column is one model as in Equation (3.3.1). The main regressor of interest is a dummy indicating the week of DST transition.

Except for drug overdosing, all estimates are negative and highly significant, mostly at the 1% level. The weekly decreases in daily admissions range from 8.3% for the all cause admission rate (column (1)) to a similar 7.5% for cardiovascular admissions (column (2)). Injuries decrease by almost 5% or about 2.7 per 1 million residents. Consistent with the medical literature (Berk, Dodd, Hallam, Berk, Gleeson, and Henry, 2008), even suicide attempts decrease by 2.76 per 100 million residents.

Next we zoom in and plot the daily estimates of Equation (3.3.1) in event study graphs. Figure 3.4.2 shows all cause admissions per 100,000 population and cardiovascular admissions per 100,000 population, respectively. Despite a conservative

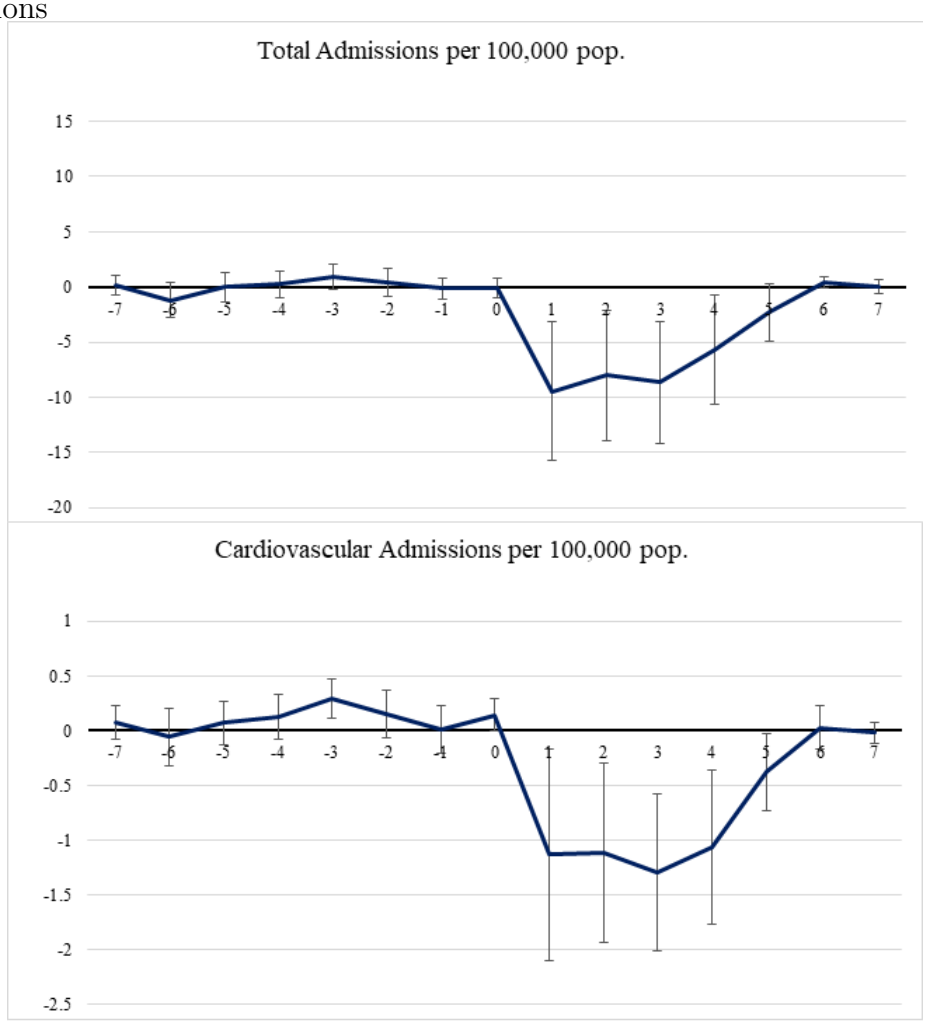
Table 3.3: Effects of Fall DST Transition on Hospitalizations by Disease Type

| | (1) All cause admissions rate | (2) Cardiovascular admissions rate | (3) Heart attack rate | (4) Injury admission rate |
|---|---|---|---|---|
| Week of Transition | -4.9556*** | -0.7195*** | -0.0882*** | -2.7121*** |
| | (1.1139) | (0.1589) | (0.02611) | (0.6869) |
| Controls | X | X | X | X |
| *Dep. var. mean* | 59.77 | 9.53 | 1.59 | 57.56 |
| $R^2$ | 0.8469 | 0.5675 | 0.1510 | 0.2067 |
| Observations | 336,604 | 336,604 | 336,604 | 336,604 |

| | (5) Metabolic admissions rate | (6) Suicide attempt rate | (7) Drug overdosing |
|---|---|---|---|
| Week of Transition | -0.1874*** | -0.0276** | -0.0044 |
| | (0.0385) | (0.0128) | (0.0055) |
| Controls | X | X | X |
| *Dep. var. mean* | 0.32 | 0.09 | 0.32 |
| $R^2$ | 0.3095 | 0.0179 | 0.0008 |
| Observations | 336,604 | 336,604 | 336,604 |

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are in parentheses and two-way clustered at the county and date level. Week of Transition is an indicator variables that equals 1 if the interview date is on the DST Sunday or one of the following 6 days. Controls include county fixed effects, vacation fixed effects, day-of-week x month fixed effects, month x year fixed effects, linear and quadratic time trends, and socioeconomic covariates. Each column is one model as in Equation (3.3.1). All admission rates are per 100,000 except for Injuries, Suicides and Drug Overdosing (per 1,000,000).

Figure 3.4.2: Effects of Fall DST Transition on Total and Cardiovascular Hospital Admissions



two-way clustering on the date and county-level, the census of hospital admissions identifies even daily effects in a very precise manner.

The two event study graphs in Figure 3.4.2 show a characteristic four-day pattern of decreases in admissions: We observe significant decreases in overall and cardiovascular admissions on days one to four after the time shift. The effect is most pronounced on the Monday after the clocks are set back, and it decreases smoothly over the next three days before it disappears on day five. The decrease for cardiovascular admissions equals about 1 avoided admission per 100,000 population for four days, or about a 10% decrease for four days.

In robustness checks, one obtains exactly the same pattern using the full sample, heart attacks and injuries, and suicide attempts. The consistency of these patterns for even heart attacks suggests that the decrease in admissions is not due to voluntary behavioral responses.

We interpret the similarity of these four-day patterns as strong support for our identification strategy. The implication is that additional sleep leads to immediate health improvements across disease groups for people who are on the margin of being hospitalized.

### 3.4.3 The Effect of Time Shift on Self-Reported Health

So far, we have seen significant reductions in hospital admissions following the time shift. However, does more sleep also make people feel better? To address this question, we again use the BRFSS. Using excellent health as the outcome, Figure 3.4.3 plots the coefficients of Equation (3.3.1). As above, all point estimates are

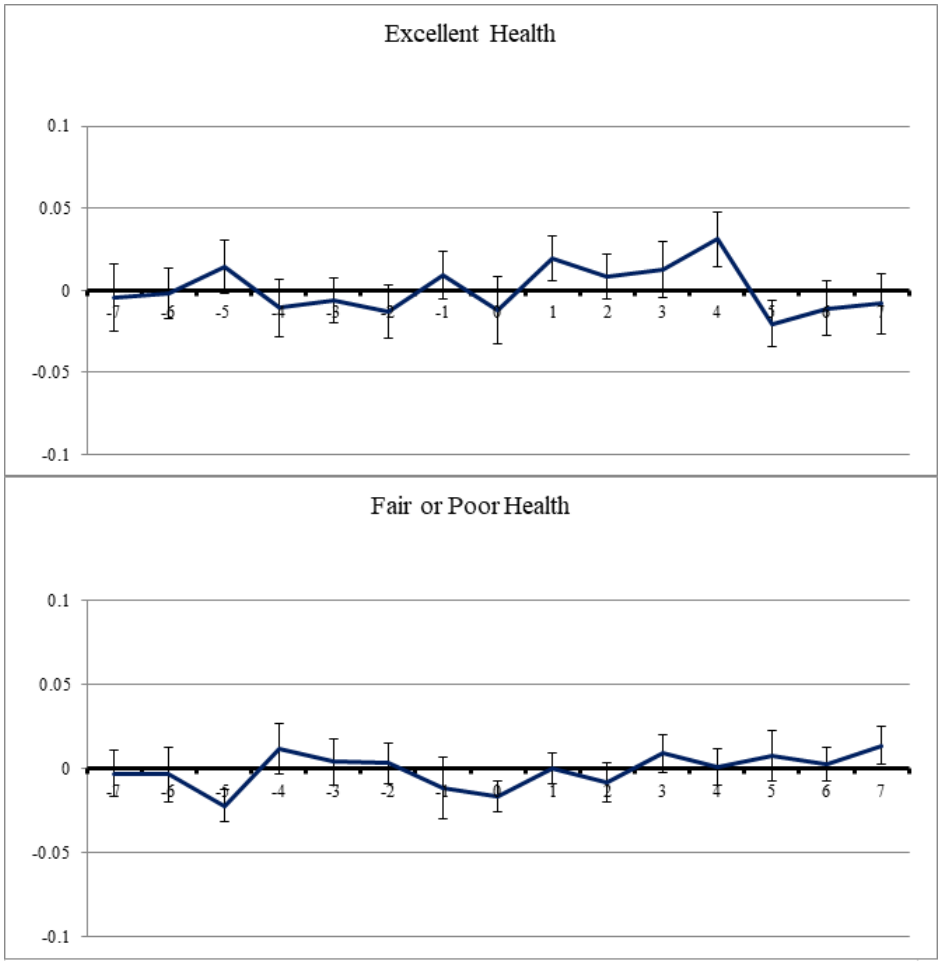
plotted along with 90% confidence intervals.

Following the time shift, the share of people who report excellent health increases by a statistically significant 1ppt on the Monday after the transition; and the effect persists until Thursday. Although the estimates are naturally noisier than the hospital admission estimates, we again observe the characteristic four-day pattern. In fact, it is strikingly similar to the pattern that we find for hospital admissions in Germany (Section 3.4.2) and our measure of tiredness (Section 3.4.1). The size of the probability-weighted coefficients would translate into about 2.5 million marginal Americans who report "excellent" instead of "very good" health for four days.

The results are robust to using all 52 weeks of the year and not weighting the regressions. The pattern also remain robust when we explore movements from SAH category three (good health) to category two (very good health).

The bottom graph of Figure 3.4.3 uses fair or poor health as outcome. Here, maybe surprisingly, we do not find significant effects. We offer a few possible explanations. First, the SAH measures are inherently noisy. For example, when we homogenize the sample, exclude people who sleep more than 8 hours on a regular basis and re-run Equation (3.3.1), we do find significant decreases in fair or poor health by 3.3ppt. A complementary explanation is that Americans are very optimistic about their health which is why only 6% report poor health. It is likely that people who "just" lack sleep do not self-categorize as being in poor health—recall that studies find that ten percent of the population are permanently sleep deprived (Knutson, Van Cauter, Rathouz, DeLeire, and Lauderdale, 2010). Finally, very sick respondents may not be able to complete the survey (e.g., because of being hospitalized),

Figure 3.4.3: Effects of Fall DST Transition on People Reporting Excellent and Poor Health



and thus may not appear in the data.

### 3.4.4 Could Alternative Mechanisms Explain the Health Effects?

Next we investigate whether alternative mechanisms could explain the health effects that we find. For example, an alternative mechanism that could theoretically produce the health benefits is the shift in ambient light from evening to morning hours. As the clocks "fall back" by one hour, sunrise and sunset both occur at earlier times. One could hypothesize that, because mornings get brighter earlier, people are more likely to exercise in the morning following the transition (and less likely to exercise in the evening). To test for this the net effect on exercising, we use a BRFSS measure on exercising and run our standard model in Equation (3.3.1). In line with Giuntella and Mazzonna (2017), we find no evidence that exercising changes as a result of the time change.

A shift in ambient light can also affect traffic accidents. However, Smith (2016) does not find that the time shift in the fall significantly affects fatalities. Moreover, traffic accidents cannot explain why hospital admissions drop sharply across a broad range of diseases, most of which are not related to accidents.

Another potential confounding factor could be crime. Doleac and Sanders (2015) show that robberies decrease in the days following the DST transition in spring (when evenings get dark later). However, they find no significant effects on crime rates in fall. Even if there was a significant robbery effect, robberies would then increase following the time shift in the fall (because it gets dark sooner), and thus

have adverse health effects, opposite the prediction of our sleep mechanism.

The fall DST transition increases the length of the Sunday from 24 to 25 hours. This may affect hospital admissions (or health survey responses) in ways unrelated to sleep. The most plausible hypothesis is that, because the day is longer, the total number of admissions will be higher, suggesting that we identify a lower bound. Moreover, this mechanism cannot explain why we find persistent health effects over four days.

## 3.5 Conclusion

This paper exploits the quasi-experimental nature of Daylight Saving Time (DST) to assess whether getting more sleep during the fall transition improves population health in the short-run. We use a large survey dataset from the US and the census of hospital admissions from Germany over one decade. Our results provide consistent and robust evidence across the two countries that health significantly improves for about four days after people gain more sleep. About 2.5 million Americans sleep significantly more following the time shift, are less likely to unintentionally fall asleep during the day, and consider themselves to be in better subjective health. Moreover, hospital data also show the same characteristic four-day drop in admissions in the days following the transition. For example, cardiovascular admissions decrease by ten admission per one million population over four days. This implies that, for people on the margin with poor health, additional sleep and rest may prevent unwanted health shocks. We also find similar patterns of reduced admissions for patients with

other diseases (which are not necessarily diagnosed on these days), but no changes in placebo tests.

The main objective of this paper is to provide evidence for the existence of a causal relationship between sleep and human capital. We do not intend to draw conclusions about the overall welfare effects of Daylight Saving Time. We would also like to point to a caveat: our reduced-form approach is well-suited for the identification of causal and immediate intent-to-treat effects, but less suited to identify long-term effects of sleep. The sleeping habits may affect mood, cognitive skills and health cumulatively over time in the long run. Alternatively, it is possible that the human body is able to adapt to (adverse) sleeping conditions. Field experiments have the power to find answers to these questions (Tepedino, Rao, Schilbach, Schofield, and Toma, 2017). More research is necessary to better understand how improvements in sleep quality may improve life quality, education and labor market outcomes as well as life expectancy.

# Bibliography

Arnold, K. C., and C. J. Flint (2017): "Vaginal Birth After Previous Cesarean Delivery," in *Obstetrics Essentials*, pp. 115–121. Springer.

Baicker, K., K. S. Buckles, and A. Chandra (2006): "Geographic variation in the appropriate use of cesarean delivery," *Health Affairs*, 25(5), w355–w367.

Banks, S., and D. F. Dinges (2007): "Behavioral and physiological consequences of sleep restriction," *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, 3(5), 519.

Berk, M., S. Dodd, K. Hallam, L. Berk, J. Gleeson, and M. Henry (2008): "Small shifts in diurnal rhythms are associated with an increase in suicide: The effect of daylight saving," *Sleep and Biological Rhythms*, 6(1), 22–25.

Betran, A., M. Torloni, J. Zhang, and A. Gülmezoglu (2016): "WHO Statement on caesarean section rates," *BJOG: An International Journal of Obstetrics & Gynaecology*, 123(5), 667–670.

Biddle, J. E., and D. S. Hamermesh (1990): "Sleep and the Allocation of Time," *Journal of Political Economy*, 98(5, Part 1), 922–943.

Billari, F. C., O. Giuntella, and L. Stella (2017): "Broadband Internet, Digital Temptations, and Sleep," *IZA DP 11050*.

Blumenthal-Barby, J., and H. Krieger (2015): "Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy," *Medical Decision Making*, pp. 539–557.

Camerer, C. F. (1989): "Does the Basketball Market Believe in the Hot Hand?," *The American Economic Review*, 79(5), 1257–1261.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011): "Robust inference with multiway clustering," *Journal of Business & Economic Statistics*, 29(2), 238–249.

Caughey, A. B., A. G. Cahill, J.-M. Guise, D. J. Rouse, A. C. of Obstetricians, Gynecologists, et al. (2014): "Safe prevention of the primary cesarean delivery," *American journal of obstetrics and gynecology*, 210(3), 179–193.

CDC (2013): "Prescription Sleep Aid Use among Adults: United States, 2005-2010," http://www.cdc.gov/nchs/data/databriefs/db127.pdf. Accessed: 2017-05-01.

Chen, D. L., T. J. Moskowitz, and K. Shue (2016): "Decision Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires," *The Quarterly Journal of Economics*, 131(3), 1181–1242.

Croson, R., and J. Sundali (2005): "The gambler's fallacy and the hot hand: Empirical data from casinos," *Journal of risk and uncertainty*, 30(3), 195–209.

Currie, J., and W. B. MacLeod (2008): "First Do No Harm? Tort Reform and Birth Outcomes," *The Quarterly Journal of Economics*, pp. 795–830.

Currie, J., and W. B. MacLeod (2017): "Diagnosing expertise: Human capital, decision making, and performance among physicians," *Journal of Labor Economics*, 35(1), 1–43.

DiSalvo, D. (2015): "How The Sleep Industry Is Making Billions From Our Sleepless Nights," https://www.forbes.com/sites/daviddisalvo/2015/08/06/how-the-sleep-industry-is-making-billions-from-your-lack-of-shuteye. Accessed: 2017-05-01.

Doleac, J. L., and N. J. Sanders (2015): "Under the cover of darkness: How ambient light influences criminal activity," *Review of Economics and Statistics*, 97(5), 1093–1103.

Dubay, L., R. Kaestner, and T. Waidmann (1999): "The impact of malpractice fears on cesarean section rates," *Journal of health economics*, 18(4), 491–522.

Emanuel, E. J., P. A. Ubel, J. B. Kessler, G. Meyer, R. W. Muller, A. S. Navathe, P. Patel, R. Pearl, M. B. Rosenthal, L. Sacks, et al. (2016): "Using behavioral economics to design physician incentives that deliver high-value carebehavioral economics, physician incentives, and high-value care," *Annals of internal medicine*, 164(2), 114–119.

Gibbons, L., J. M. Belizán, J. A. Lauer, A. P. Betrán, M. Merialdi, F. Althabe, et al. (2010): "The global numbers and costs of additionally

needed and unnecessary caesarean sections performed per year: overuse as a barrier to universal coverage," *World health report*, 30, 1–31.

Gibson, M., and J. Shrader (2018): "Time use and productivity: The wage returns to sleep," *Review of Economics and Statistics*.

Gilovich, T., R. Vallone, and A. Tversky (1985): "The hot hand in basketball: On the misperception of random sequences," *Cognitive psychology*, 17(3), 295–314.

Giuntella, O., and F. Mazzonna (2017): "Sunset time and the economic effects of social jetlag. Evidence from US time zone borders," *Unpublished manuscript, Department of Economics, University of Pittsburgh*.

Green, B., and J. Zwiebel (2017): "The hot-hand fallacy: Cognitive mistakes or equilibrium adjustments? Evidence from Major League Baseball," *Management Science*.

Gruber, J., J. Kim, and D. Mayzlin (1999): "Physician Fees and Procedure Intensity: the Case of Cesarean Delivery," *Journal of Health Economics*, 18, 473–490.

Gruber, J., and M. Owings (1996): "Physician Financial Incentives and Cesarean Section Delivery," *RAND Journal of Economics*, 27, 99–123.

Guryan, J., and M. S. Kearney (2008): "Gambling at lucky stores: Empirical evidence from state lottery sales," *The American Economic Review*, 98(1), 458–473.

Haelle, T. (2016): "Your Biggest C-Section Risk May Be Your Hospital," *Consumer Reports, April*, 13.

Hamermesh, D. S., C. K. Myers, and M. L. Pocock (2008): "Cues for timing and coordination: Latitude, Letterman, and longitude," *Journal of Labor Economics*, 26(2), 223–246.

Hillman, D. R., A. S. Murphy, R. Antic, and L. Pezzullo (2006): "The economic cost of sleep disorders," *Sleep*, 29(3), 299–305.

Johnson, E. J., and D. Goldstein (2003): "Medicine. Do defaults save lives?," *Science*, 302(5649), 1338–1339.

Johnson, E. M., and M. M. Rehavi (2016): "Physicians treating physicians: Information and incentives in childbirth," *American Economic Journal: Economic Policy*, 8(1), 115–141.

Khullar, D., D. A. Chokshi, R. Kocher, A. Reddy, K. Basu, P. H. Conway, and R. Rajkumar (2015): "Behavioral economics and physician compensation—promise and challenges," *New England Journal of Medicine*, 372(24), 2281–2283.

Killgore, W. D. (2010): "Effects of sleep deprivation on cognition," in *Progress in brain research*, vol. 185, pp. 105–129. Elsevier.

Knutson, K. L., E. Van Cauter, P. J. Rathouz, T. DeLeire, and D. S. Lauderdale (2010): "Trends in the prevalence of short sleepers in the USA: 1975–2006," *Sleep*, 33(1), 37–45.

Martin, J., B. Hamilton, M. Osterman, A. Driscoll, and T. Mathews (2017): "Births: Final Data for 2015," Discussion paper, Center for Disease Control and Prevention.

McGinnis, J. M., L. Stuckhardt, R. Saunders, M. Smith, et al. (2013): *Best care at lower cost: the path to continuously learning health care in America*. National Academies Press.

Miller, J. B., and A. Sanjurjo (2014): "A cold shower for the hot hand fallacy," *IGIER Working Paper*.

——— (2015): "Surprised by the gambler's and hot hand fallacies? A truth in the law of small numbers," *IGIER Working Paper*.

Molina, G., T. G. Weiser, S. R. Lipsitz, M. M. Esquivel, T. Uribe-Leitz, T. Azad, N. Shah, K. Semrau, W. R. Berry, A. A. Gawande, et al. (2015): "Relationship between cesarean delivery rate and maternal and neonatal mortality," *Jama*, 314(21), 2263–2270.

Moore, P. J., N. E. Adler, D. R. Williams, and J. S. Jackson (2002): "Socioeconomic status and health: the role of sleep," *Psychosomatic medicine*, 64(2), 337–344.

Mullington, J. M., M. Haack, M. Toth, J. M. Serrador, and H. K. Meier-Ewert (2009): "Cardiovascular, inflammatory, and metabolic consequences of sleep deprivation," *Progress in cardiovascular diseases*, 51(4), 294–302.

Persistent Market Research (2015): "Global Sleep Aids Market to Account for US80.8Bn by 2020," http://www.persistencemarketresearch.com/mediarelease/sleep-aids-market.asp. Accessed: 2017-05-01.

Raab, M., B. Gula, and G. Gigerenzer (2012): "The hot hand exists in volleyball and is used for allocation decisions.," *Journal of Experimental Psychology: Applied*, 18(1), 81.

Rabin, M., and D. Vayanos (2010): "The gambler's and hot-hand fallacies: Theory and applications," *The Review of Economic Studies*, 77(2), 730–778.

Shurtz, I. (2013): "The impact of medical errors on physician behavior: Evidence from malpractice litigation," *Journal of Health Economics*, 32(2), 331–340.

Sirovich, B. E., S. Woloshin, and L. M. Schwartz (2011): "Too little? Too much? Primary care physicians' views on US health care: a brief report," *Archives of internal medicine*, 171(17), 1582–1585.

Smith, A. C. (2016): "Spring forward at your own risk: daylight saving time and fatal vehicle crashes," *American Economic Journal: Applied Economics*, 8(2), 65–91.

Suetens, S., C. B. Galbo-Jørgensen, and J.-R. Tyran (2015): "Predicting Lotto Numbers: A Natural Experiment on the Gambler's Fallacy and the Hot-Hand Fallacy," *Journal of the European Economic Association*.

Taheri, S., L. Lin, D. Austin, T. Young, and E. Mignot (2004): "Short sleep duration is associated with reduced leptin, elevated ghrelin, and increased body mass index," *PLoS medicine*, 1(3), e62.

Tepedino, P., G. Rao, F. Schilbach, H. Schofield, and M. Toma (2017): "Pre-analysis plan outline for Sleepless in Chennai: the economic effects of sleep deprivation among the poor," Discussion paper, Working paper, MIT, https://www.socialscienceregistry.org/docs/analysisplan/1365/document.

Witte, D. R., D. Grobbee, M. L. Bots, and A. W. Hoes (2005): "A meta-analysis of excess cardiac mortality on Monday," *European journal of epidemiology*, 20(5), 401–406.